UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

PREDICCIÓN DE LA CORRUPCIÓN VÍA RED NEURONAL

Autor: Roberto Sánchez Fernández
Tutor: David Renato Domínguez Carreta

Febrero 2017

PREDICCIÓN DE LA CORRUPCIÓN VÍA RED NEURONAL

AUTOR: Roberto Sánchez Fernández
TUTOR: David Renato Domínguez Carreta

Dpto. Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Febrero de 2017

Resumen (castellano)

Este Trabajo Fin de Grado comprende un exhaustivo examen sobre la corrupción en los países, realizándose un estudio sobre su evolución durante de un periodo de tiempo para así poder realizar una predicción de su desarrollo futuro.

Por ello se ha realizado un proceso de recopilación de datos relativos al barómetro general de corrupción de los años 2004, 2005, 2006, 2010 y 2013, recopilados por Transparencia Internacional.

Para acometer la tarea de predecir el desarrollo futuro de los niveles de corrupción se emplean técnicas de machine-learning tales como vecinos próximos, el perceptrón multicapa o los árboles de decisión. Adicionalmente se ha valorado la calidad de las predicciones mediante la convergencia de los exponentes de Lyapunov y el cálculo de la tasa de error en la fase de testeo.

Como paso previo a estas predicciones se ha realizado un proceso de estudio de los datos recogidos, incluyendo el empleo de *clusters*, el cálculo de la entropía de cada uno de los distintos campos de que se disponía o el establecimiento de enlaces entre los distintos países en función de las distancias euclidianas de los valores de sus campos para cada año.

Estas pruebas y predicciones se han realizado empleando la herramienta Weka. Además, se han desarrollado herramientas propias para cubrir tareas necesarias para el desarrollo del Trabajo, tales como una serie de programas que convierten conjuntos de datos en un fichero de entrada para Weka, así como un programa que enlaza con Weka a través de línea de comandos para realizar las predicciones.

Abstract (English)

This Bachelor Thesis provides an exhaustive study about corruption, being the subject of study its evolution through a time period in order to predict its future development.

In order to do so, a process of data gathering has been performed, acquiring data gathered by International Transparency relative to the general corruption barometer from years 2004, 2005, 2006, 2010 and 2013.

To accomplish the task of predicting the future developments of the corruption level in a certain country, a set of machine learning algorithms, such as the multilayered perceptron, the nearest neighbors and decision trees, have been used. Additionally, a process of quality assurance has been performed on the aforementioned algorithms, using the Lyapunov exponent and the value of the error percentage during test to do so.

Before the aforementioned predictions a study of the available data has been performed, using clusters as well as the calculation of the entropy of the various fields available in the data set and the search for links between different countries based on the euclidean distances between the various fields corresponding to each country in a specific year.

These experiments and predictions have been done using the Weka program. Additionally, a set of programs have been coded to do some certain tasks, such as transforming raw data into an arff file that Weka can understand or a program that calls Weka classifiers through a command line interface, required during the process of making the Thesis.

Palabras clave (castellano)

Corrupción, *Machine Learning*, *cluster*, Weka, Red Neuronal

Keywords (inglés)

Corruption, Machine Learning, cluster, Weka, Neural Network

Agradecimientos

A mi tutor, David, por su inestimable ayuda al realizar éste trabajo.

A mi familia, por hacer posible que estudie ésta carrera.

A mis amigos, que, con sus luces y sus sombras, contribuyen a que todo esto merezca la pena.

A todos ellos:
Gracias.

INDICE DE CONTENIDOS

Contenido Página

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Uno de los problemas más acuciantes en las sociedades modernas es la corrupción. La visibilidad de este problema se ha visto aumentada por la crisis económica que ha diezmado las economías mundiales desde finales de la década pasada. Un ejemplo de ésta visibilidad es que, en la encuesta del CIS de diciembre de 2015, la corrupción aparece listada como la segunda causa de preocupación para los españoles con un 38.8% de menciones, sólo por detrás del paro (79.8%) y por encima de otras cuestiones como la política (14.8%), la situación económica (24.4), la sanidad o la educación (11.9 y 9.7%, respectivamente). [1]

Otro ejemplo de la visibilidad conseguida por la corrupción es que, durante la campaña electoral previa a las elecciones generales del 20 diciembre de 2015, fue uno de los temas de mayor relevancia en los programas políticos de los distintos partidos que concurrían a las mismas.

Adicionalmente, si miramos los informes publicados por la organización Transparencia Internacional, que, entre otros estudios, publica una clasificación de los países según su índice de percepción de corrupción con valores entre 0 y 100 (cuánto más alto sea el valor, menor es el nivel de corrupción), podemos ver, para el informe de 2015, a España situada en el puesto 37 con una puntuación de 56, muy lejos del primer puesto, que ocupa Dinamarca, con una puntuación de 90 y por detrás de países como, por ejemplo, Emiratos Árabes Unidos o Uruguay.[2]

Desde un punto de vista macroeconómico es posible establecer una relación entre el nivel de corrupción de un país tiene relación directa con la situación económica del mismo (nuevamente si nos dirigimos a los informes de Transparencia Internacional y los comparamos con el puesto que ocupa cada país ordenado por renta per cápita se ve la relación muy claramente, especialmente para el caso de países netamente corruptos).

Por todo esto se puede concluir que estudiar el desarrollo de la corrupción es una buena forma de predecir el desarrollo económico futuro, proveyendo de valiosa información a inversores y empresas que busquen expandir su mercado.

1.2 Objetivos

En este Trabajo de Fin de Grado se toman como objetivos primordiales tanto realizar un estudio de la corrupción a nivel mundial, como una predicción de su desarrollo futuro.

Primeramente, es necesario realizar un estudio de los datos para obtener una serie de ideas acerca de cómo proceder más adelante. Para esto lo que se pretende es emplear la API de Weka, probando distintos clasificadores para intentar inferir relaciones y características de los datos. Después se pretende emplear técnicas de *clustering*, también usando la API de Weka, para realizar un estudio de a) qué países se catalogarían como corruptos y cuáles no. (empleando 2 *clusters*) y b) qué países corruptos están en la senda de ser clasificados como

no corruptos y vice versa, es decir, cuáles de los no corruptos pueden estar en retroceso hacia la corrupción (empleando 3 *clusters*).

No se contempla el incluir datos relativos a la posición geográfica de los distintos países para la realización del agrupamiento, pues se desea que la clasificación se haga exclusivamente en términos de los parámetros que tengan relación con la corrupción. No obstante, sí que se considera que puede ser un valor añadido a la hora de realizar la predicción.

Una vez realizado el estudio de los datos se pasará a realizar una predicción, utilizando para ello el algoritmo *backpropagation*, Para ello se llevará a cabo un proceso de optimización del aprendizaje de la red previo a la clasificación. Por último, se estudiará, además la calidad de éstas predicciones mediante el cálculo del exponente de Lyapunov.

1.3 Metodología Empleada

Para acometer el desarrollo de los distintos programas necesarios para este Trabajo se ha adoptado un modelo de desarrollo iterativo en cascada. De este modo se procedió a realizar un diseño personalizado de cada uno de los programas requeridos para, posteriormente, pasar a codificar el proyecto y ponerlo a prueba. Si durante las pruebas se encontraba algún fallo se procedía a retornar a la fase de codificación o de diseño para hacer las modificaciones necesarias. A continuación, se detalla programa a programa la metodología empleada:

1. Convertidor de ficheros de bgcs a arff: Este programa está desarrollado en C. En su desarrollo, primeramente, se procedió a determinar el modo en que se tratarían los datos no presentes debido a que en los documentos usados para obtener los datos no se presentaba una estructura uniforme, por lo que fue necesario inferir datos a partir de los resultados de otros años. Posteriormente se diseñaron lo tipos abstractos de datos que almacenarían los valores del bgc para cada par país/año.

2. Generador de arrfs de predicción: Este programa está desarrollado en C y genera los ficheros arrf que requieren los programas Weka para hacer la predicción. Para ello introduce todos los países con sus regiones para los años del período 2016-2020, además de generar un arff de entrenamiento con los datos de los años 2004,2005,2006 y 2010 y otro de test con los de 2013, por limitaciones de Weka, se asigna una clase a los valores (véase sección 3.1.2).

3. Calculador de coeficiente de Lyapunov: Programa escrito en C. Lee todo el conjunto de datos de entrada de un fichero con países ya clasificados y obtiene el coeficiente de Lyapunov.

4. Programa que muestra mapas del mundo con distintos códigos de color: Aplicación web escrita en HTML y JavaScript que usa la API Geochart de Google. Toma como parámetros un año y un gráfico (Barómetro Global de Corrupción, Clasificación o Clúster) y dibuja un mapamundi con un código de colores según el valor requerido, que aplica a cada país según corresponda.

5. Programas generadores de ficheros de clusterización: Ambos programas están escritos en C. Los dos generan un arff con la cabecera y datos pertinentes para los conjuntos de valores de BGC (sin incluir país o año). La diferencia estriba en que uno

de ellos genera un único fichero con todos los datos y el otro genera un fichero por año con los datos de ese año.

6. Programas que usan la API Java Weka: Conjunto de programas que implementan el entrenamiento y testeo de conjuntos de datos con distintos algoritmos (*Backpropagation*, KNN, *Naïve Bayes*…) variando distintos parámetros para optimizar el error, permitiendo también realizar la clasificación de un nuevo conjunto de datos, así como la clusterización usando el algoritmo K-Means.

1.4 Exploración de conceptos teóricos utilizados

1.4.1 Entropía

En física, la entropía es un concepto utilizado para medir el desorden de un sistema, por ejemplo, las moléculas de un gas, que al estar en expansión tienden a desordenarse, dando por tanto una entropía mayor que en un sólido, donde se encuentran muy unidas y, por tanto, más ordenadas. [3]

Éste concepto fue reutilizado más adelante en el marco de la teoría de la información desarrollada por *Claude Shannon* y *Warren Weaver* [4], siendo adaptado como medición de la incertidumbre y está representada por la siguiente ecuación:

$$H(p)=- \Sigma(p_i * \log(p_i))$$

Donde la base del logaritmo puede ser la que se quiera. Para el presente trabajo, a la hora de calcular entropías, se ha optado por base 2. [5]

En ésta fórmula, p_i representa la probabilidad de que un valor se encuentre en un intervalo de valores. Para ésta prueba se han generado 2 intervalos (representando corrupción o no), tomando como valor de separación el medio del rango de valores posibles (cómo va de 1 a 5, se ha elegido 3).

1.4.2 Exponente de Lyapunov

1. Este conjunto de operaciones tiene como objetivo desarrollar una medición de la velocidad en que varían los valores de un sistema a lo largo de un período de tiempo. Este valor se obtiene como resultado del cálculo:

   $$L=lim\ (1/t * ln\ (D_T/D_0))$$

   Donde t indica la diferencia de tiempo, en valor absoluto, entre el momento X y el de referencia (en el caso del conjunto de datos es el año 2004, por ser el primero del que hay datos) y D_T/D_0 corresponde a la diferencia (en valor absoluto) entre los valores del año de referencia con respecto al actual.

A partir de este cálculo se presentan dos variantes principales con respecto al valor de L.

- Si L es menor o igual que 0, se puede extraer que el sistema es estable.

- Si L es, por el contrario, mayor que 0, se puede extraer que el sistema es caótico.

De aquí se puede inferir la tendencia futura en base a los valores que va tomando L en los distintos períodos. Si L va aumentando, el sistema está desordenándose, y si va decreciendo, éste irá ordenándose. Por tanto, es sencillo comprender la relación producida entre este valor y la calidad de la predicción, si el sistema tiende al orden, la predicción será más precisa que si el sistema tiene tendencia al caos. [6]

1.4.3 Algoritmos de Machine-Learning

1.4.3.1 Vecinos Próximos

Este algoritmo establece como clasificación la clase mayoritaria entre los K vectores de entrenamiento con menor distancia de entre el conjunto de vectores de que se dispone, estableciendo la distancia mediante métodos como la distancia euclídea (calculando la raíz cuadrada de la suma de las diferencias entre los pares de campos elevadas al cuadrado) o la distancia Manhattan (sumando 1 a la distancia por cada par de campos que sean distintos entre los dos vectores).

A modo de ejemplo, es posible establecer un paralelismo entre este sistema y el empleado para resolver problemas a partir de "ejercicios tipo". También se puede emplear este método para establecer una prueba de la validez de los datos a la hora de clasificar (a partir de un conjunto de campos lo suficientemente extenso), si conjuntos de datos muy similares no tienen como resultado tasas de error bajas, se podría afirmar que esos campos no son muy significativos a la hora de establecer la clase del dato.

1.4.3.2 Perceptrón Multicapa

Es un tipo de red neuronal que, a diferencia del perceptrón mono capa, puede resolver problemas no separables linealmente, empleando para ello el sistema de *backpropagation*, La salida se consigue mediante la propagación de una capa a otra de la activación obtenida por una entrada, aplicando para ello una función de activación que realiza una combinación lineal de las activaciones de las neuronas de una capa con los pesos correspondientes a la conexión de esas neuronas con las de la capa anterior.

La diferencia fundamental entre el perceptrón simple y el perceptrón multicapa se da en el algoritmo de aprendizaje. Este, en sus primeras etapas es igual que el *feedforward* del perceptrón simple, pero, una vez llega a la(s) neurona(s) de salida, realiza un proceso de *feedback* del resultado con respecto al valor esperado a las capas anteriores, que provoca un proceso de ajuste del aprendizaje.

1.4.3.3 Naïve Bayes

Consiste en realizar una clasificación bayesiana asumiendo como cierto que todos los atributos son independientes entre sí, esto quiere decir, se asume que el valor de un atributo no está derivado de ningún otro atributo presente en el vector de clasificación (por ejemplo, asumir que el que haya humedad es independiente del día de la semana que sea a la hora de predecir si va a llover o no). La clasificación final es aquella que resulte tener la mayor probabilidad, siendo esta obtenida mediante la multiplicación de todas las probabilidades dada cada clase.

Como consecuencia lógica de la asunción previamente enunciada, la tasa de error es inversamente proporcional al grado de independencia entre los atributos.

Existe la posibilidad de incluir la llamada corrección de Laplace, que asume que hay al menos un caso de todas las posibles variantes clase/atributo, de modo que, si no se tiene una combinación concreta en el conjunto de entrenamiento, no se asume que la combinación al completo es improbable. Aplicar esta corrección suele mejorar ligeramente el porcentaje de error obtenido en clasificación.

Además, igual que ocurre con todos los clasificadores bayesianos, éste clasificador sólo puede funcionar con clases con valores discretos.

1.4.3.1 Redes Bayesianas

Consiste en un modelo de clasificador bayesiano que permite configurar las relaciones de dependencia entre distintos campos, al contrario que el algoritmo bayesiano normal (que asume dependencia completa) y *Naïve Bayes* (que asume independencia completa), resultando mejor a la hora de clasificar conjuntos de datos complejos, siempre que se realice un correcto estudio previo de las dependencias.

1.4.3.2 Árboles de decisión

Este tipo de algoritmos se caracterizan por realizar la clasificación en como resultado de una serie de respuestas binarias a preguntas. Existen dos modelos de árboles de decisión, cuya única diferencia es el tipo de pregunta que realizan. Por un lado, el modelo ID3, que realiza preguntas del tipo ¿valor == x?, lo que lo convierte en un buen método para clasificar datos discretos, y por otro, el modelo C4.5, que realiza preguntas del tipo ¿valor > x?, con lo que se adapta mejor a los valores continuos.

Ni el método ID3 ni el método C4.5 sirven para realizar predicciones sobre clases continuas, por razones obvias (el número de hojas del árbol sería infinito, con lo que en términos de uso de memoria sería exageradamente ineficiente).

1.4.4 Clustering

Se conoce como *clustering* a la tarea de agrupar distintos vectores de datos (por ejemplo, los niveles de corrupción en distintos sectores de la vida pública) en una serie de grupos (llamados *clusters*). Existen múltiples algoritmos diferentes para realizar esta tarea, que dan lugar a múltiples definiciones de lo que es un *cluster*, siendo la más común aquella que lo limita a grupo de vectores cuyas distancias son mínimas.

Un ejemplo de algoritmo que emplea esta aproximación es *K-Means*. El funcionamiento de este algoritmo es el siguiente:

Dado un número X de *clusters*, se generan X centroides aleatorios (bien obteniendo valores aleatorios, bien eligiendo X vectores aleatorios del conjunto de datos) y se procede a asignar, según un criterio de distancia mínima, cada vector del conjunto a uno de los *clusters*. Una vez se han asignado todos los vectores, se procede a calcular un nuevo centroide (típicamente con la media de los valores de los vectores del *cluster*) para cada *cluster* y se repite la

asignación según distancia. Este proceso se repite hasta que se consigue la estabilización de los centroides, aunque también se suele incluir un límite de iteraciones, como medida de prevención ante la posibilidad de caer en un bucle infinito.

1.5 Organización de la memoria

El presente documento se estructura en torno a cinco grandes apartados siendo estos los siguientes:

- Introducción: En este apartado se explica brevemente tanto la motivación que lleva al desarrollo de este trabajo como los objetivos que se pretenden alcanzar en la realización de los experimentos a que se llevarán a cabo a lo largo del mismo. Adicionalmente, se ofrece una breve exposición de la metodología utilizada para el desarrollo de los distintos programas que se han empleado para poder completar el trabajo, así como de los conceptos teóricos utilizados a lo largo del trabajo.

- Estado del Arte: Esta sección consta de tres apartados principales:
    - El primero constituye un estudio sobre la corrupción en general.
    - El segundo es un estudio a grandes rasgos sobre el programa Weka, es decir su historia, interfaz, sus distintos apartados, etc.
    - En el tercer y último punto se realiza una revisión histórica del *machine learning*, haciendo especial énfasis en los hitos que h.

- Metodología: A lo largo de éste capítulo, se explicarán los procedimientos adoptados a la hora de realizar el pre procesamiento de los datos, así como se detallarán los distintos experimentos (así como las conclusiones obtenidas de los mismos) realizados para obtener una mayor comprensión de los datos empleados, exponiéndose información obtenida mediante *machine learning*m, las distintas clusterizaciones empleadas, el establecimiento de enlaces entre países y los resultados de los distintos cálculos de entropías.

- Predicción: En este apartado se explican los procedimientos empleados para realizar la predicción futura del desarrollo de los niveles de corrupción, así como los resultados obtenidos una vez completada. Se incluye además un estudio de la calidad de la misma a partir del cálculo de los exponentes de Lyapunov.

- Conclusiones y Trabajo Futuro: En este apartado se realiza una breve recapitulación, a modo de cierre, de los objetivos del Trabajo; así como de los resultados que se han obtenido en las distintas pruebas realizadas. Por último, se procede a detallar posibles formas de ampliar el presente trabajo en el futuro.

2 Estado del arte

2.1 Corrupción

"El poder tiende a corromper y el poder absoluto corrompe absolutamente"

Lord Acton (1834-1902)

Según la RAE, la corrupción (en su definición más acorde con el propósito de este trabajo) es "una práctica consistente en la utilización en organizaciones, especialmente en las públicas, de funciones y medios de aquellos en provecho, económico o de otra índole, de sus gestores". [11]

Uno de los principales problemas de la sociedad moderna es la corrupción, especialmente la corrupción en las instituciones públicas. Es por ello que, desde principios del presente siglo, la Organización No Gubernamental Transparencia Internacional publica, de manera anual, una serie de estudios sobre la corrupción, desglosados por país, principalmente en Índice de Percepción de Corrupción y el Barómetro Global de Corrupción, de modo que se pueda ir evaluando la evolución de la misma en los distintos países. [7]

2.2 Weka

2.2.1 Introducción

Weka es un programa desarrollado por la Universidad de Waikato (Nueva Zelanda) cuyo desarrollo comenzó en 1993. Se trata de un programa de software libre, publicado bajo la Licencia Pública General de GNU.

Originalmente escrito en C, en 1997 se tomó la decisión de reescribirlo entero en Java, lenguaje en el que ha continuado su desarrollo hasta la actualidad. Como consecuencia de esto, es un programa altamente portable, lo que, añadido a la posibilidad de emplear una interfaz gráfica, otorga a este programa una gran accesibilidad, tanto desde el punto de vista del acceso al programa y de la facilidad de uso del mismo. [8]

Weka proporciona una suite de algoritmos de aprendizaje automático, incluyendo algoritmos como Naïve Bayes, Vecinos Próximos, la clasificación a priori, regresión logística o distintos tipos de redes neuronales, como el perceptrón multicapa.

Weka utiliza como formato de entrada el formato arff (Attribute Relation File Format). Este formato consiste en incluir los datos, separando cada columna por una coma, habiendo incluido previamente una cabecera del fichero con meta información, principalmente acerca de los datos que conforman cada columna, siendo estos o bien NUMERIC o bien una lista de los valores que puede tomar, separando cada valor con una coma e incluyendo el conjunto al completo entre llaves Un ejemplo sería éste: {Finlandia, Suecia, Noruega}.

2.2.2 Modo de empleo

Weka presenta varios modos de empleo. Primeramente, se puede dividir entre el modo gráfico y el modo texto (*simple cli*).

Figura 1: Interfaz Inicial de Weka

Un ejemplo de ejecución en el modo texto sería el siguiente:

java <classname> <args> [> file]

De entre los distintos modos gráficos que se aprecian en la Figura 1, sólo se va a estudiar el modo *Explorer*, que ha sido el empleado en este Trabajo.

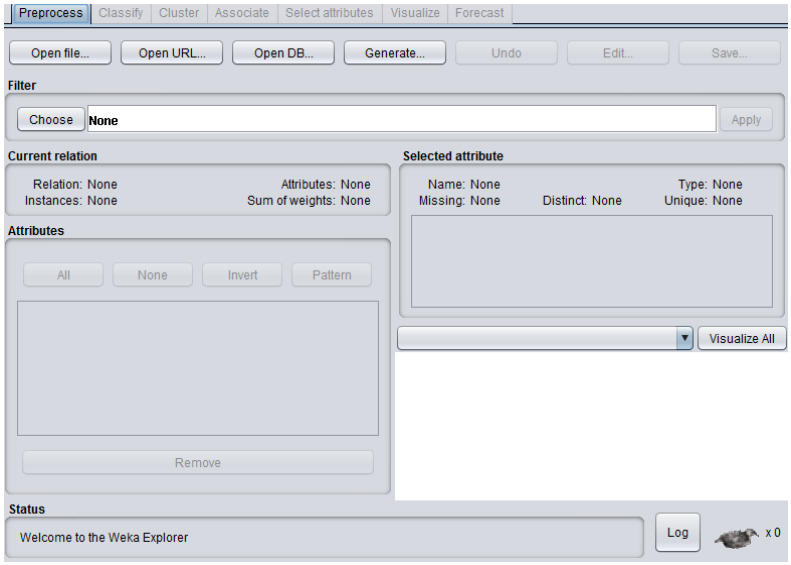

Figura 2: Interfaz explorer de Weka

En este modo, se nos presentan varias pestañas, de las cuales nos interesan cuatro:

En la pestaña *Preprocess* es donde introduciremos el fichero de entrenamiento que vayamos a usar. Una vez leído el fichero se nos mostrarán datos sobre el mismo, como el número de instancias (datos) o la distribución de las clases, en forma de histograma.

En la pestaña *Classify* seleccionaremos el algoritmo de aprendizaje (o los algoritmos, en caso de que usemos el clasificador por votación) y el método de test, entre *cross validation* (pudiendo personalizar el número de particiones) y *percentage split* (pudiendo personalizar el porcentaje).

La pestaña *Cluster* es en base igual a la pestaña *Classify* salvando alguna diferencia (no permite elegir *cross validation*, por ejemplo).

También se ofrece una completa API en Java con la que se pueden elaborar programas que amplíen las capacidades de Weka para, por ejemplo, realizar una predicción automatizada para todos los países en un período de varios años y procesar los datos en ficheros de salida similares a los que contenían los datos originales.

En los programas desarrollados se han empleado los siguientes métodos de dicha API:

- *buildClassifier*: Este método de la clase *Classifier* se encarga de cargar un array de *Instances* como conjunto de entrenamiento del clasificador y realizar el entrenamiento. Ëstas se han cargado directamente desde un *BufferedReader* que abre el fichero arff que se quiere utilizar.
- *setOptions*: Este método también corresponde a la clase *Classifier*, recibe como argumento un objeto del tipo *Options*, en el caso de los programas realizado, éste era una instancia de *splitOptions* conformada a partir de un *String*.
- *evaluateModel*: Método de la clase *Evaluation*, recibe un objeto de la clase *Classifier* ya entrenado y uno de la clase *Instance*, donde se encuentran los datos a testear. Este método clasifica cada una de las instancias del conjunto de test y genera el porcentaje de error, obtenible llamando a la función *errorRate()* sobre el mismo objeto.
- *classifyInstance*: Este método también corresponde a la clase *Classifier*, recibe como argumento un objeto de la clase *Instance* y retorna el valor de la clase asignada.
- *setNumClusters*: Método de la clase *SimpleKMeans*. Establece número de clusters a usar.
- *buildClusterer*: Método de la clase *SimpleKMeans* equivalente a *buildClassifier*.

2.3 Machine Learning

El *machine learning* o aprendizaje automático es una rama de la inteligencia artificial que se centra en conseguir reproducir el mecanismo de aprendizaje presente en los seres vivos, principalmente en el ser humano, en forma de algoritmo, para que pueda ser entendido y replicado de manera autónoma por un ordenador.

Actualmente, algunos de los usos de éstos algoritmos son el diagnóstico de enfermedades a partir de los síntomas, la secuenciación de cadenas de ADN, el reconocimiento del lenguaje hablado y escrito o la detección de fraude con tarjetas de crédito.

La historia del aprendizaje automático se remonta a los años 50 con la proposición teórica de la Máquina de Aprendizaje de Turing (*Turing A. 1950*) y la aparición de los primeros modelos de red neuronal, siendo los principales hitos la implementación del aprendizaje hebbiano por Farley y Clark (*Farley 1954*) , la invención del perceptrón por Frank Rosenblatt en 1957 y la invención de ADALINE por Widrow en 1960, momento a partir del cual se

suele fijar el comienzo de la llamada edad dorada de las redes neuronales, que terminaría a finales de dicha década cuando Marvin Minsky y Seymour Papert demostraron (*Minsky, M. & S. Papert 1969*) que ambos modelos de red neuronal y, en general, todos los sistemas *feedforward*, presentaban el mismo problema: eran incapaces de resolver problemas no linealmente dependientes. Éste problema se resolvería (*Werbos 1975*) con la introducción del algoritmo Backpropagation (véase sección 1.4.3.2) al abandonar el sistema de aprendizaje exclusivamente *feedforward*.

A partir de este momento, se producen principalmente dos grandes innovaciones, la aparición de las redes neuronales recurrentes (RNN), como pueden ser las redes de Hopfield (1982, aunque ya habían sido propuestos por Little en 1974) o los mapas auto organizativos de Kohonen (*Kohonen, Teuvo 1982*), y la utilización de arquitecturas hardware paralelizadas, el conocido como *deep learning*.

Además del desarrollo de las redes neuronales cabe destacar como eventos importantes la invención del algoritmo de vecinos próximos (KNN véase sección 1.4.3.1) en 1951 por parte de Fix y Hodges (*Fix, E. & J.L. Hodges 1951*), el desarrollo de los algoritmos de árbol de decisión (véase sección 1.4.3.4), como ID3 (*Quinlan 1986*) o su sucesor C4.5 (*Quinlan 1993*) y de los algoritmos bayesianos como Naïve Bayes o las redes Bayesianas que serían empleados (como alternativa a otras técnicas como la aplicación de reglas lógicas) en los sistemas expertos como Dendral (Edward Feigenbaum 1965) o Mycin (desarrollado a principios de los años 1970 en la Universidad de Stanford por Edward Shortlife).[9][10]

3 Metodología

3.1 Preprocesamiento de los datos

3.1.1 Barómetro Global de Corrupción

A continuación, se muestra una imagen de cómo se encontraban los datos originalmente:

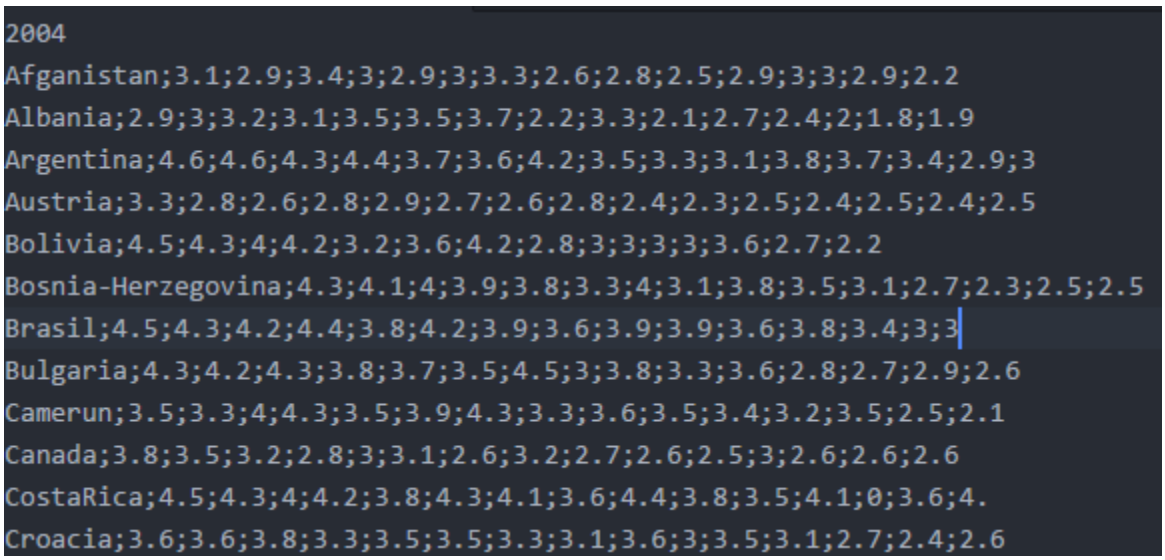

Figura 3: Fichero de datos original

Como podemos ver, el fichero tiene una estructura en bloques con la siguiente forma: primero se incluye un año para posteriormente incluirse los valores del barómetro global de corrupción para cada país, separados por un punto y coma. Cada "vector" consta de los siguientes campos: nombre de país y valor BGC de los distintos sectores (a saber, partidos políticos, parlamento, justicia, fuerzas de seguridad del estado, sector empresarial, hacienda, aduana, medios de comunicación, sector sanitario, sector educativo, servicios públicos, Fuerzas Armadas, ONGs, instituciones religiosas y las oficinas de registro del estado).

En el programa de pre procesamiento se procedía a cargar en un TAD apropiado la información relativa a todos los países, al tiempo que se les asignaba una región geográfica (esto último se hacía buscando el país en un *array* que contenía los pares país – región para todos los países del conjunto).

Las regiones contempladas son: Europa Continental, Europa Mediterránea, Europa del Norte, Norteamérica, Latinoamérica, Magreb, África Subsahariana, Oriente Medio, Sudeste asiático y Oceanía.

```
typedef struct _pais {
    char name[PAIS];
    char region[PAIS];
    anyovalor valores[ANYOS];
} pais_s;
```

Figura 4: TAD de país-año

```
typedef struct _anyovalor {
    int anyo;
    float partidopolitico;
    float parlamento;
    float judicial;
    float policia;
    float empresas;
    float impuestos;
    float aduanas;
    float medios;
    float sanidad;
    float educacion;
    float serviciospublicos;
    float ejercito;
    float ongs;
    float religion;
    float registro;
} anyovalor;
```

Figura 5: TAD de año

Una vez leído el fichero se procedía a realizar un barrido de todos los países con objeto de encontrar años "en blanco" – debido a que ni los conjuntos de países ni los sectores estudiados son constantes para los distintos años- y se realizaba una asignación de esos valores calculando un valor aleatorio entre el máximo y mínimo que presentara ese país en el resto de años para cada campo.

Finalmente se procedía a imprimir los nuevos datos en forma de fichero, con la siguiente estructura:

```
2010,Afganistan,AsiOrM,2.90,3.20,3.40,3.20,3.10,3.00,3.30,2.80,2.89,2.90,3.10,2.90,3.10,2.70,2.20
2013,Afganistan,AsiOrM,3.00,3.10,3.70,2.90,3.00,3.00,3.30,2.40,2.90,2.90,3.30,2.40,2.90,2.30,2.20
2004,Albania,EurCent,2.90,3.00,3.20,3.10,3.50,3.50,3.70,2.20,3.30,2.10,2.70,2.40,2.00,1.80,1.90
2005,Albania,EurCent,3.61,3.09,4.10,3.68,2.78,3.09,3.54,2.55,3.60,2.65,3.47,2.64,2.86,2.05,1.90
2006,Albania,EurCent,3.20,3.20,3.10,3.80,3.80,2.50,3.40,4.10,2.80,2.70,3.20,3.60,3.00,2.30,1.90
2010,Albania,EurCent,4.00,3.89,4.28,3.27,3.26,3.35,3.45,2.82,3.83,2.46,2.80,3.49,2.50,1.91,1.90
2013,Albania,EurCent,4.10,3.90,4.30,3.70,2.70,2.84,3.52,2.90,4.30,4.00,3.50,2.90,2.30,1.80,1.90
```

Figura 6: Fichero de salida

3.1.2 Asignación de Clases

El objetivo principal de añadir estas clases es facilitar el proceso de aprendizaje y predicción, ya que Weka no puede clasificar más de un valor por "vector" así como ampliar el número de métodos de *machine learning* que se pueden emplear para el estudio, ya que algunos son incompatibles con clasificación continua (véase sección 1.4), razón por la cual se ha optado por tomar clases abstractas a partir de los datos (no limitándose a, por ejemplo, la mediana de entre los valores de cada vector).

Bajo esta premisa se ha optado por establecer tres clases, *Clean*, *Transition* y *Corrupt*.

Para establecer cada clase se ha optado por calcular la media y varianza promedio de todos los vectores de datos y, a partir de esos dos datos, establecer las clases según los siguientes criterios:

- Si el valor promedio de un vector era menor que el promedio de todos menos la mitad de la varianza, se le asignaba la clase *Clean*.
- Si el valor promedio de un vector era mayor que el promedio de todos más la mitad de la varianza, se le asignaba la clase *Corrupt*.
- El resto eran clasificados como *Transition*.

```
void mediasyVarianzas (pais_s * paises, float * media, float * varianza){

  int i, j;
  float mediaParcial=0;
  float *f;
  int count = 0;
  *media=*varianza = 0;

/*Media*/
  for(i=0; i <124!= 0 ;i++){
    for(j=0; paises[i].valores[j].anyo < 2014; j++){
      mediaParcial=0;
      for(f=&paises[i].valores[j].partidopolitico; f<=&paises[i].valores[j].registro; f++){
        mediaParcial+=*f/15;
      }
      *media += mediaParcial;
      if(mediaParcial != 0){
        count++;
      }
    }
  }
  *media/=count;

  /*Varianza*/
    for(i=0; paises[i].valores[2].partidopolitico != 0 ;i++){
      for(j=0; paises[i].valores[j].anyo < 2014; j++){
        mediaParcial=0;
        for(f=&paises[i].valores[j].partidopolitico; f<=&paises[i].valores[j].registro; f++){
          mediaParcial+=*f/15;
        }
        *varianza += sqrt((mediaParcial-*media) * (mediaParcial-*media));
      }
    }
    *varianza/=count;
}
```

Figura 7: Cálculo de Media y Varianza

.

3.2 Estudio de los datos con algoritmos de machine learning

Primero se probó a generar un fichero en el que los datos de entrada al sistema de clasificación (un perceptrón multicapa) son los valores de corrupción en los distintos sectores para una serie de años (2004,2005, 2006, 2010 y 2013), siendo el objetivo clasificar el país al que pertenecen esos datos. Para hacer el estudio más interesante, se ha ido variando el número de iteraciones de entrenamiento para ver la evolución de la tasa de fallo. A continuación, se muestran los resultados:

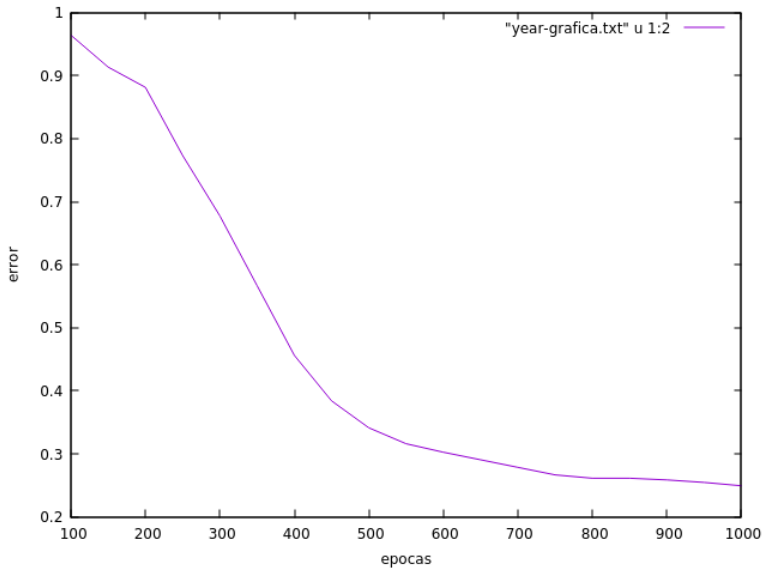

Figura 8: Evolución de la tasa de error sin regiones

Como se puede ver, la tasa de fallo se reduce rápidamente para estabilizarse en torno al 25% a partir de las 800 iteraciones de entrenamiento. En principio es una tasa de fallo bastante buena teniendo en cuenta lo limitado de los datos de entrada con respecto al número de posibles países de salida (124).

Posteriormente se realizó la misma prueba, pero añadiendo como entrada la región a la que pertenece el país que se quiere clasificar. Con esto además se podría inferir si, en principio, hay una relación entre la posición geográfica de un país y su nivel de corrupción interna. Nuevamente se realizó la prueba variando el número de iteraciones de entrenamiento para ver la evolución de la tasa de fallo. A continuación, se muestran los resultados:

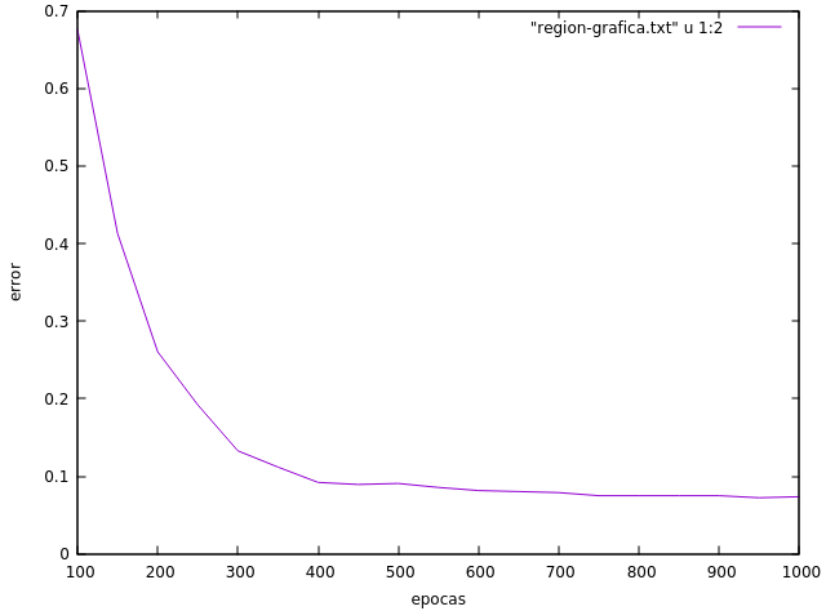

Figura 9: Evolución de la tasa de error con regiones

Nuevamente podemos ver cómo el error baja rápidamente hasta situarse por debajo del 10% a partir de las 500 iteraciones. A partir de éstos resultados podemos establecer una primera hipótesis:

Si el error al clasificar un país conociendo su región es pequeño, quiere decir que, dentro de cada región, un país tiende a ser suficientemente diferente del resto como para que sea bastante identificable. (Teniendo 124 países y 10 regiones se clasifica de promedio 1 país mal en cada región), por lo que se puede plantear que no existe una relación directa entre el origen geográfico de un país y su nivel de corrupción, al menos a un nivel tan bajo de organización por grupos.

No obstante, siempre cabe la posibilidad de que dentro de una región los países sean bastante identificables pero que entre regiones se dé el mismo suceso, implicando que los países de cada región se mueven en un rango relativamente único (podría haber solapamientos) de valores).

Para resolver esta duda se ha planteado un tercer caso que consistiría en predecir la región a la que pertenece un conjunto de datos conformado únicamente por el año y los diversos niveles de corrupción por sectores. Esto no es más que una modificación de la primera prueba, cambiando el objetivo a clasificar del país a la región. A continuación, se muestran los resultados de la evolución según el número de iteraciones de entrenamiento empleadas:

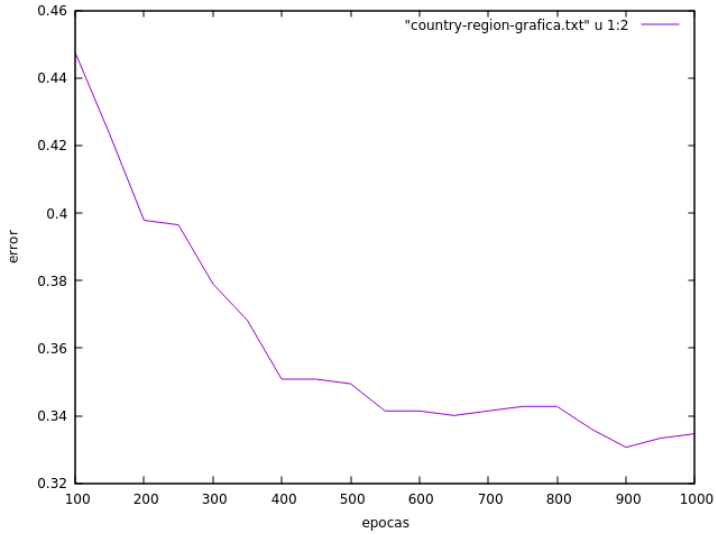

Figura 10: Evolución de la tasa de error para regiones

Como podemos ver, igual que antes, el error baja conforme va aumentando el número de épocas, aunque en un rango mucho más reducido que el de los casos anteriores y no llegando a alcanzar el nivel de error de ninguno de ellos. En base a estos resultados se puede afirmar que la hipótesis planteada previamente debe ser cierta, pues estos valores de error contradicen la posibilidad alternativa presentada de que fueran las regiones las que tuvieran niveles de corrupción característicos, ya que un 33% de fallo cuando hay 10 posibles clasificaciones es un número muy elevado de error.

Cabe pensar que esto se podía haber refutado con los resultados de la primera prueba, en la que se intentaba clasificar los países sin información de la región a la que pertenecían, dando como resultado un nivel de error no mucho mayor que cuando se introducía la región, pero siempre podría haberse debido al solapamiento antes mencionado de los valores entre unas regiones y otras (por ejemplo, las contiguas, ya que es lógico suponer que habrá regiones de transición, que seguramente se podrían detectar con una división más detallada de las regiones). No obstante, esto no quiere decir que los niveles de corrupción no puedan ser similares en los países que constituyen una región, simplemente que son bastante identificables dentro de la misma.

Por último, se presenta una comparativa entre las tres pruebas para contextualizar mejor los resultados de cada una:

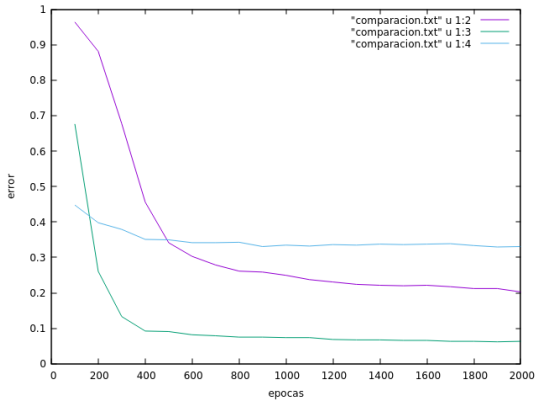

Figura 11: Comparación de los tres experimentos realizados

3.3 Clusterización

Como siguiente método de estudio de los datos utilizados se ha elegido realizar un proceso de clusterización de los mismos mediante el algoritmo K-Means. Para ello se ha optado por usar en las distintas ejecuciones dos y tres *clusters*, respectivamente, con objeto de separar los países en corruptos y no corruptos, añadiendo una tercera categoría (algo así como países en transición de un estado a otro) en el segundo caso. Para la clusterización se ha trabajado, además, con varios conjuntos de datos. Primeramente, se ha trabajado con un conjunto de datos distinto para cada año, de modo que obtenemos para cada año qué países entran en cada clasificación. Además, se ha optado por utilizar un conjunto de datos que aglutine todos los datos de la serie histórica ya que se entiende que así se puede reflejar mejor la evolución de los países a lo largo de los distintos años de que se dispone, más allá de si en un año en concreto un país X es considerado corrupto o no.

Ambos conjuntos de datos están conformados por filas de datos como los mostrados a continuación, con el valor de corrupción de cada sector separado por una coma:

```
3.10,2.90,3.40,3.00,2.90,3.00,3.30,2.60,2.80,2.50,2.90,3.00,3.00,2.90,2.20
2.90,3.00,3.20,3.10,3.50,3.50,3.70,2.20,3.30,2.10,2.70,2.40,2.00,1.80,1.90
4.60,4.60,4.30,4.40,3.70,3.60,4.20,3.50,3.30,3.10,3.80,3.70,3.40,2.90,3.00
```

Figura 12: Datos de entrada para el agrupamiento

A continuación, se muestran los valores para el primer año y el último de la serie histórica (2004 y 2013, respectivamente). Primero se muestra con los conjuntos de datos separados por años y después el de todos los datos juntos. El conjunto de resultados completo se muestra en el Anexo A.

Conjuntos anuales:

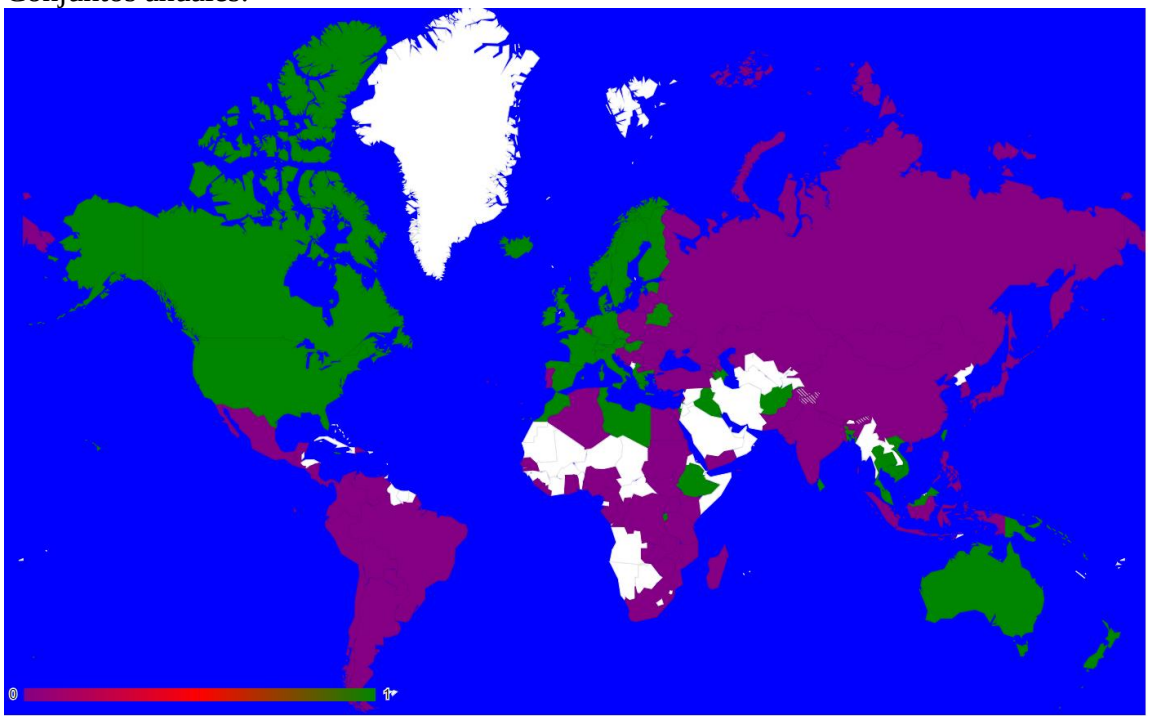

Figura 13: Agrupamiento en 2 clusters 2004

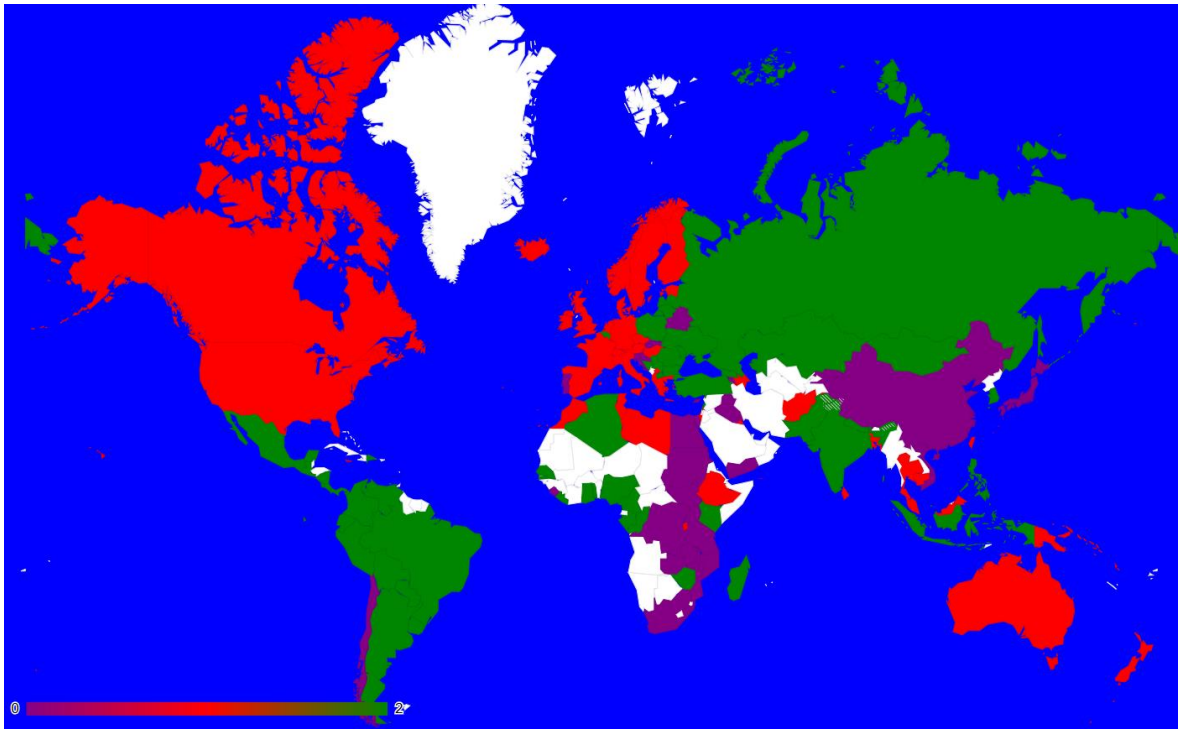

Figura 14: Agrupamiento en 3 clusters 2004

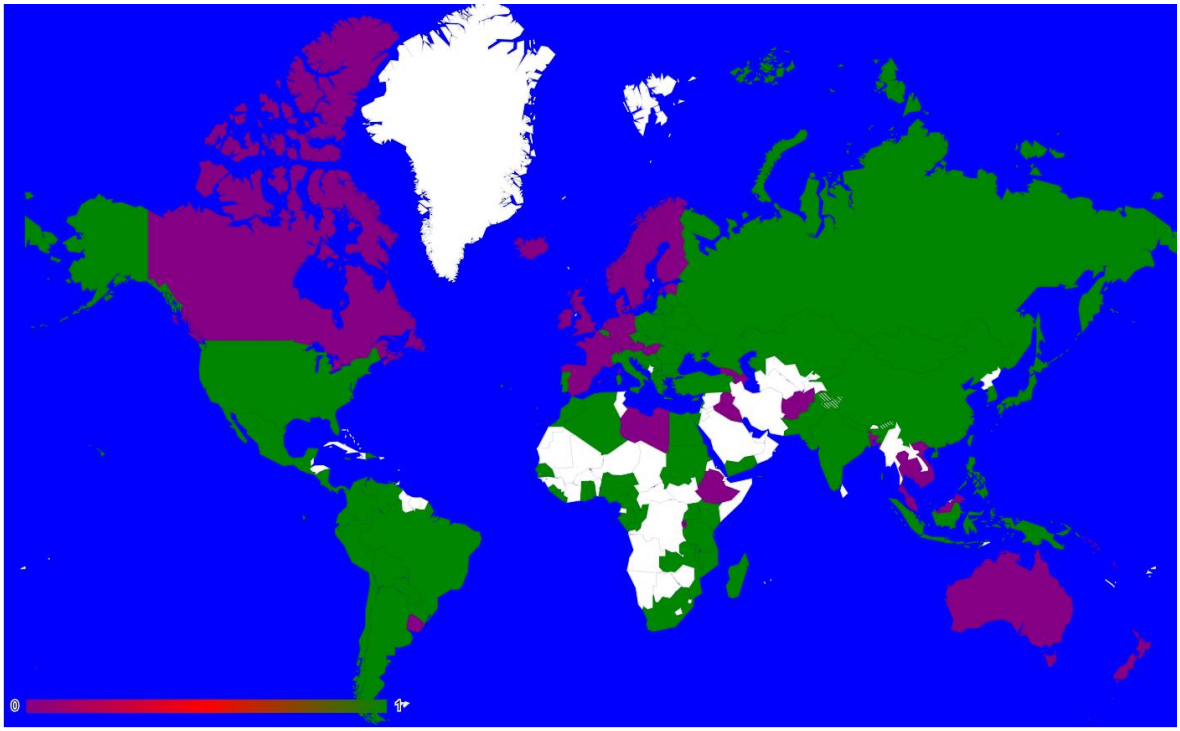

Figura 15: Agrupamiento en 2 clusters 2013

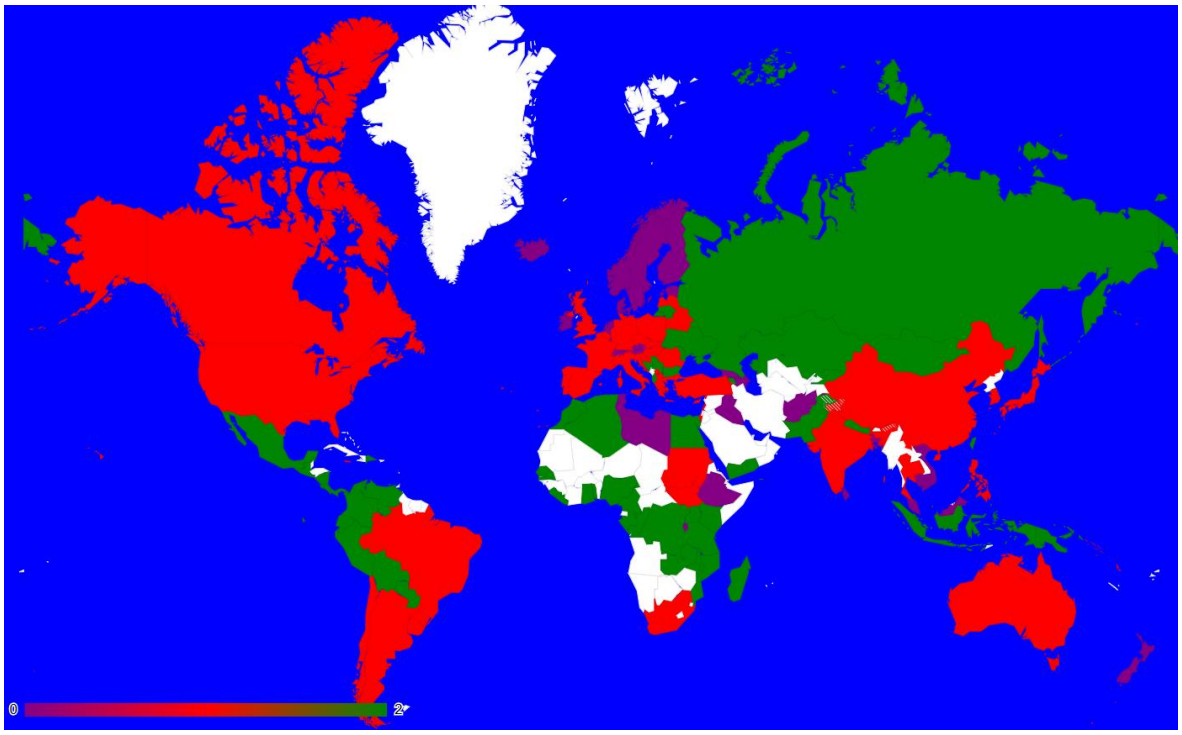

Figura 16: Agrupamiento en 3 clusters 2013

Primeramente, es necesario aclarar que la representación en colores varía como consecuencia de ser distintas clusterizaciones, unas veces los países de un tipo son asignados a un *cluster* y otras, a otro.

Del estudio de las agrupaciones realizadas por el algoritmo se pueden sacar una serie de conclusiones:

- La primera es que entre 2004 y 2013 se produjo un pequeño retroceso en cuanto a los niveles de corrupción, focalizado sobre todo en Occidente, con una caída en Estados Unidos y la Europa continental, especialmente la zona más oriental del Mediterráneo.

- La segunda es que, a nivel de clúster, existe una relación geográfica entre los países con niveles de corrupción similares, así como se puede notar la influencia de los grandes países (basta con fijarse en los cúmulos formados alrededor de potencias regionales como China, Rusia, Alemania o Brasil).

- La tercera es la estacionalidad general de los conjuntos, especialmente si dividimos los países en el eje corrupto-no corrupto ya que, como podemos ver, la mayoría de cambios se dan, precisamente en países que en 2004 eran considerados no corruptos y que en 2013 si lo son, mientras que el cambio contrario es mucho más infrecuente.

Conjunto mixto:

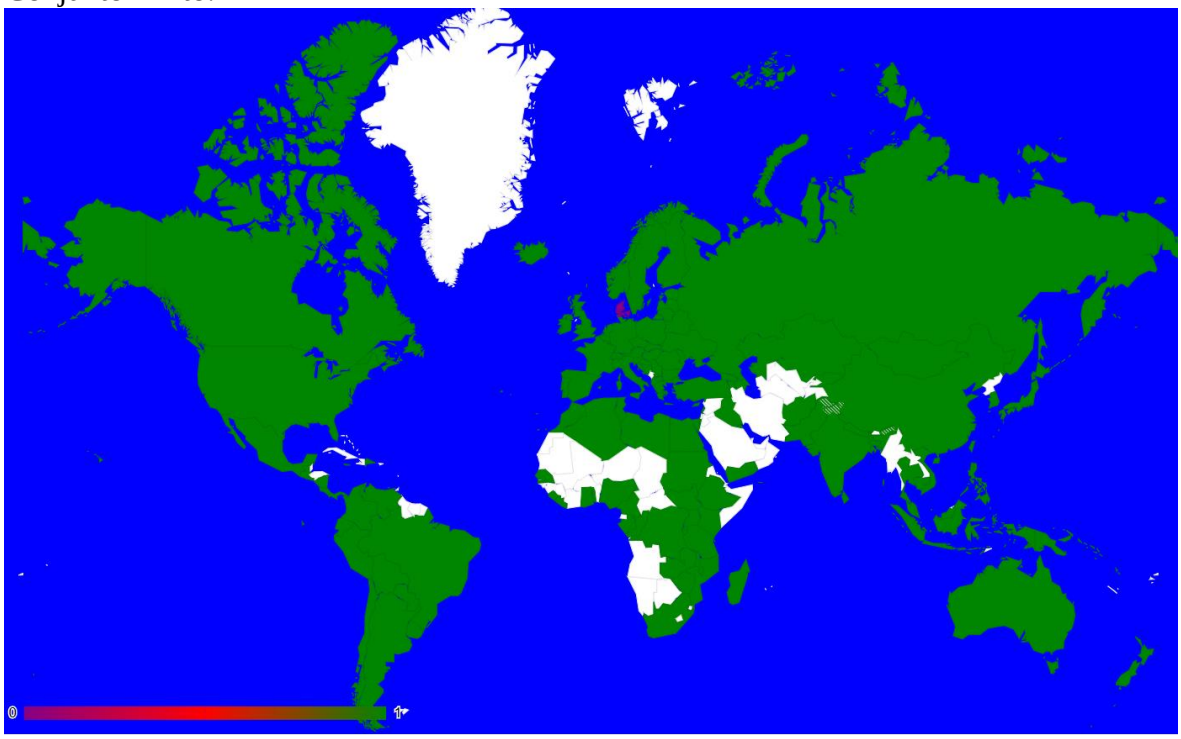

Figura 17: Agrupamiento 2 clústers 2004

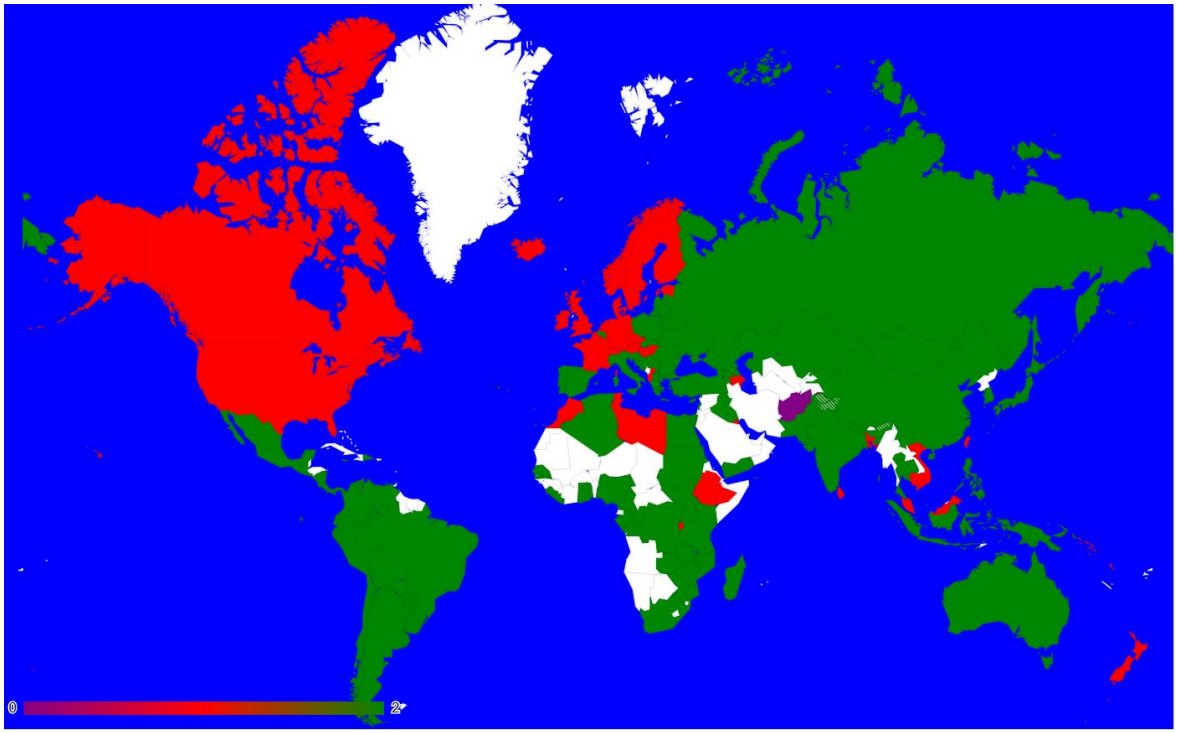

Figura 18: Agrupamiento 3 clústers 2004

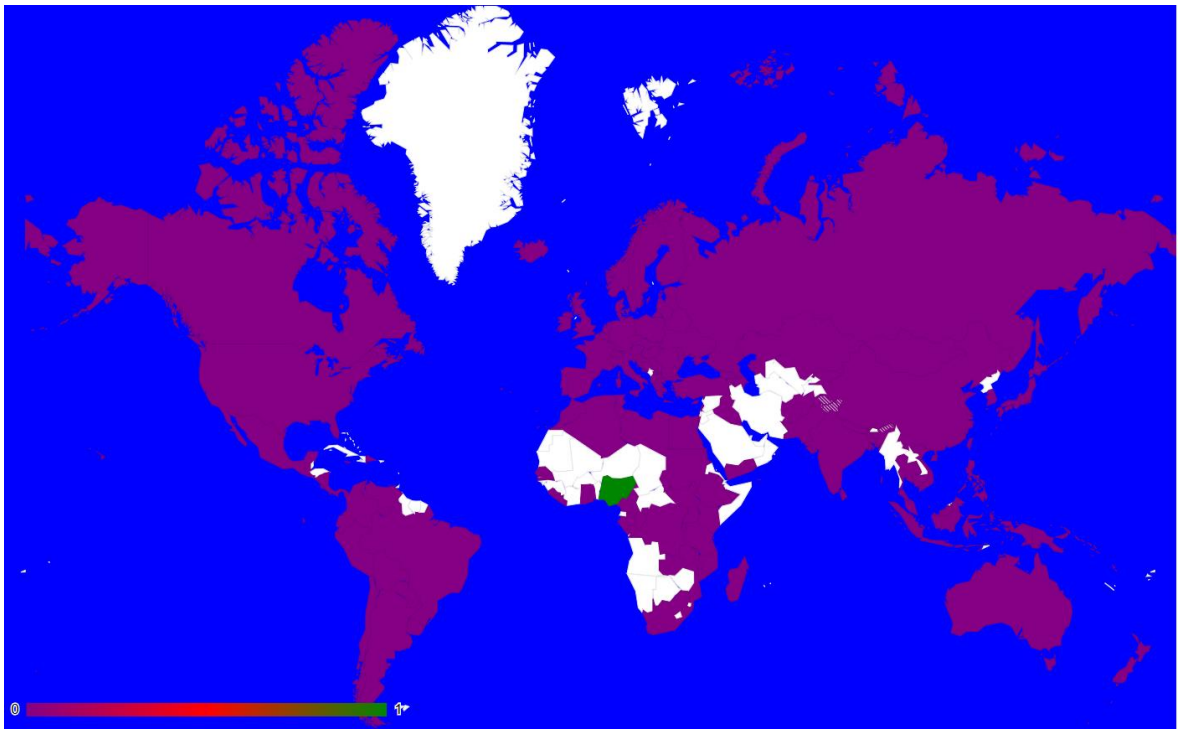

Figura 19: Agrupamiento 2 clústers 2013

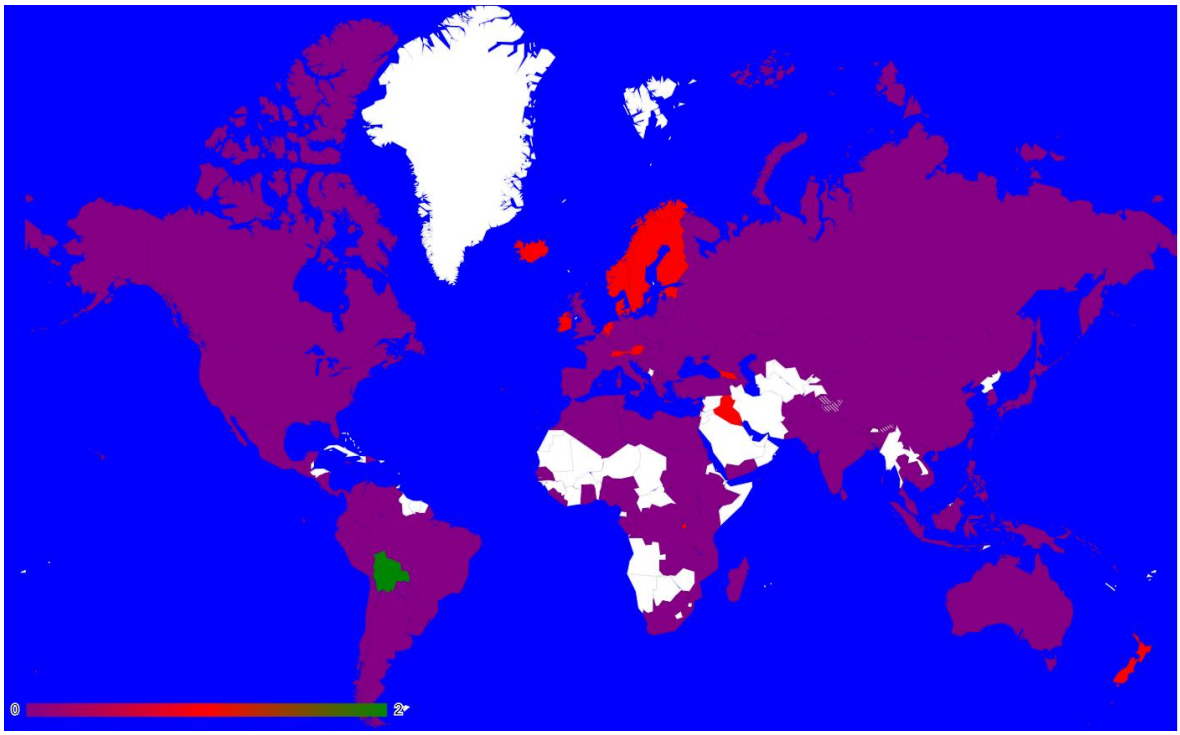

Figura 20: Agrupamiento 3 clústers 2013

En éstas gráficas, los colores se mantienen porque sólo se ha realizado una clusterización

Revisándolas, se pueden ver obtener nuevamente varias conclusiones:

- La primera es que se ha producido una mejoría en términos generales de los niveles de corrupción, aunque nuevamente se ve un retroceso en los países del primer mundo, igual que para la prueba anterior.
- También se confirma que existe una variabilidad general en el concepto de "país corrupto", pues no se aplica el mismo criterio para determinar la corrupción de un país en 2004 que en 2013.

3.4 Establecimiento de enlaces:

Como siguiente prueba, se procedió a realizar un estudio de los enlaces entre regiones dentro de un mismo año. Para ello se usó un sistema de establecimiento de enlaces entre dos países basado en la diferencia entre las medias de los respectivos valores de corrupción por sector. Para la realización de esta prueba se utilizó como valor límite de la diferencia entre medias de 0.1. Si el valor era menor este, se procedía a añadir a las estructuras correspondientes dentro de un *array* los correspondientes enlaces al contador de la región del otro país. Esto se hizo estandarizando el orden de regiones dentro de ambos *arrays* (el de regiones y el de contadores).

```
typedef struct _node {
    char region[TAM];
    int Links[REGIONES];

} node_s;
```

Figura 21: Estructura empleada para contar enlaces

A continuación se muestran los resultados obtenidos para el año 2004. Dada la similitud general de los datos anuales, así como la gran cantidad de ellos que hay, se ha optado por incluir sólo los datos de este año y hacer referencia a los datos de otros años cuando se produzcan diferencias significativas. El resto de datos se pueden ver en el Anexo B.

region	*EurCent*	*EurNor*	*EurSur*	*AmerSur*	*AmerNor*	*AsiOrM*	*IndoChina*	*Oceania*	*AfrSubSah*	*AfrSah*
EurCent	*76*	*57*	*55*	*55*	*17*	*34*	*56*	*10*	*33*	*34*
EurNor	*57*	*26*	*29*	*20*	*6*	*14*	*30*	*3*	*10*	*13*
EurSur	*55*	*29*	*50*	*66*	*10*	*30*	*50*	*3*	*32*	*38*
AmerSur	*55*	*20*	*66*	*140*	*3*	*40*	*72*	*0*	*46*	*59*
AmerNor	*17*	*6*	*10*	*3*	*2*	*5*	*10*	*4*	*3*	*1*
AsiOrM	*34*	*14*	*30*	*40*	*5*	*28*	*30*	*3*	*33*	*24*

IndoChina	*56*	*30*	*50*	*72*	*10*	*30*	*42*	*7*	*30*	*32*
Oceania	*10*	*3*	*3*	*0*	*4*	*3*	*7*	*2*	*0*	*0*
AfrSub Sah	*33*	*10*	*32*	*46*	*3*	*33*	*30*	*0*	*48*	*43*
AfrSah	*34*	*13*	*38*	*59*	*1*	*24*	*32*	*0*	*43*	*24*

Tabla 1: Enlaces del año 2004

Para poder realizar un análisis más detallado de los datos, a continuación se muestran una serie de gráficos que permiten ver la relación entre los datos con mayor profundidad:

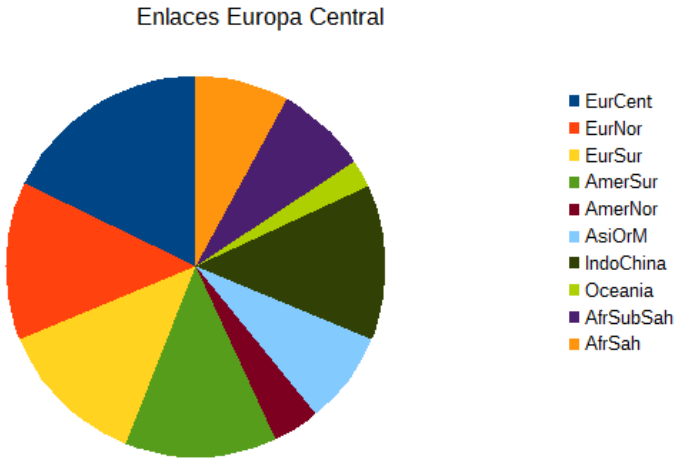

Figura 22: Enlaces Europa Central

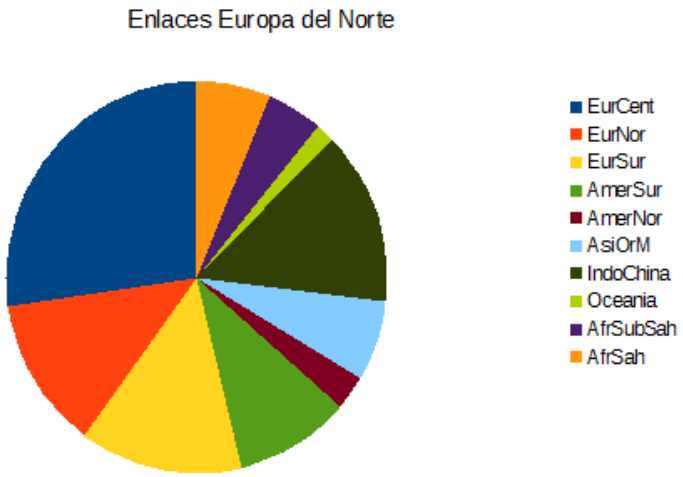

Figura 23: Enlaces Europa del Norte

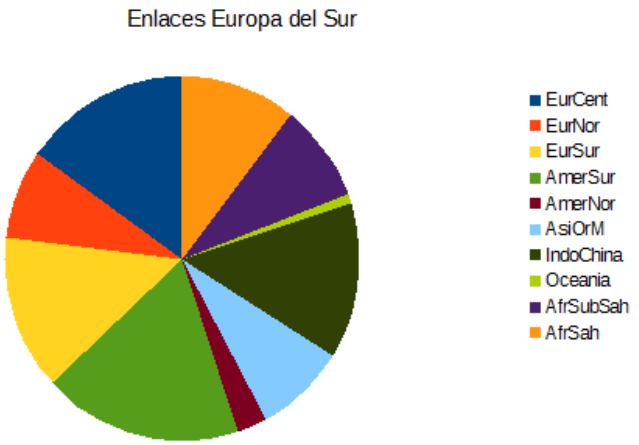

Figura 24: Enlaces Europa del Sur

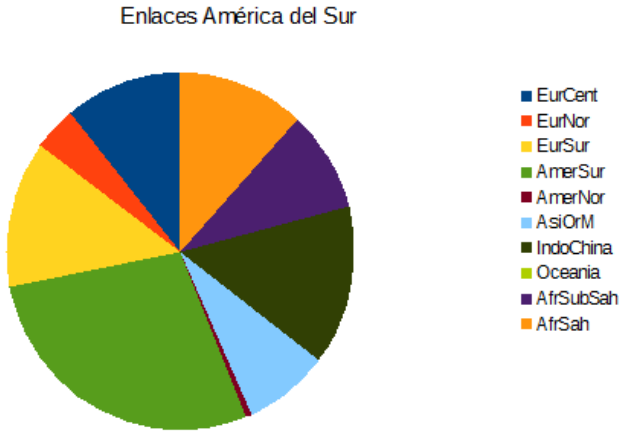

Figura 25: Enlaces América del Sur

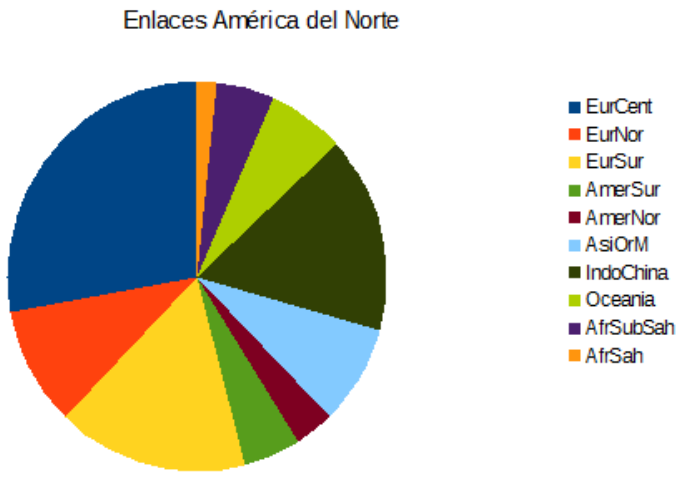

Figura 26:Enlaces América del Norte

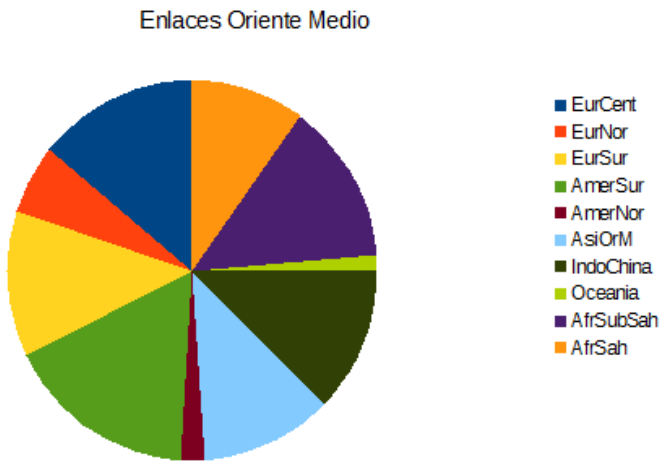

Figura 27: Enlaces Oriente Medio

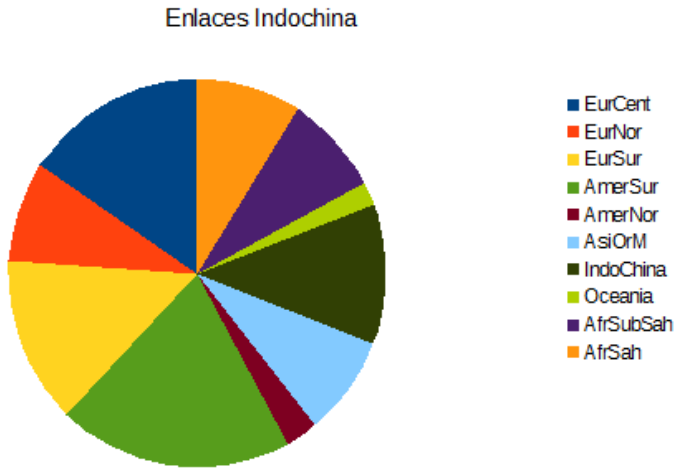

Figura 28: Enlaces Indochina

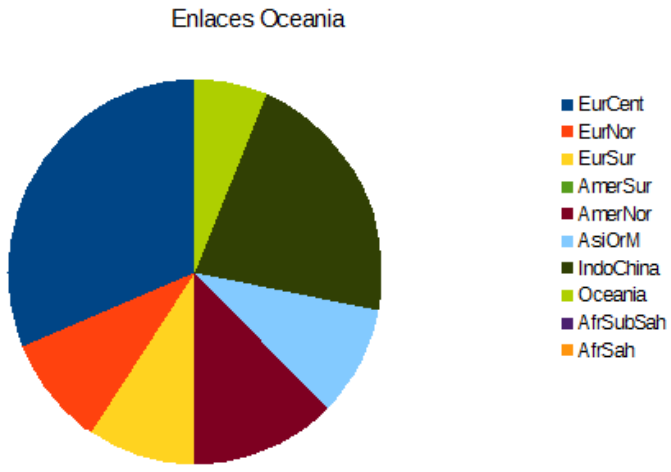

Figura 29: Enlaces Oceanía

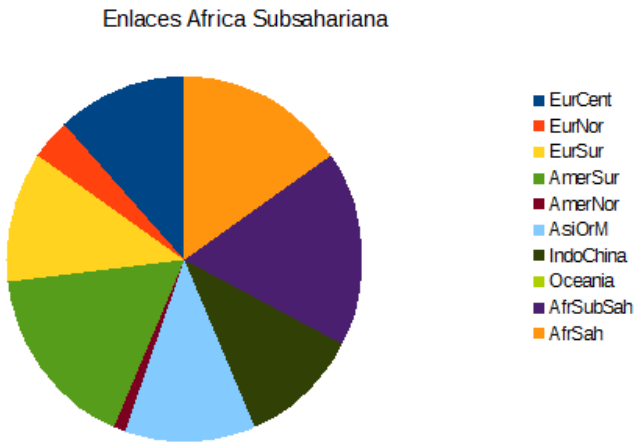

Figura 30: Enlaces África Subsahariana

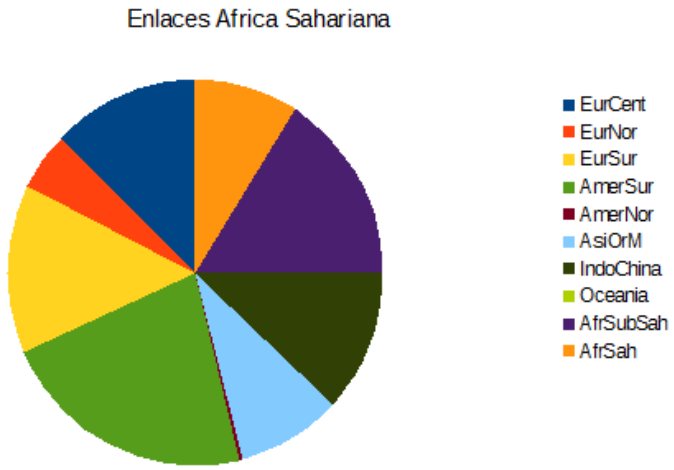

Figura 31: Enlaces África Sahariana

Una característica general que se puede extraer de los datos es que hay un mayor nivel de conectividad con regiones contiguas (aunque en, por ejemplo, América, esto no se dé) o de similar nivel de desarrollo económico-social, también se puede ver una cierta relación entre los antiguos territorios coloniales y las regiones a las que pertenecen sus antes colonizadores.

Podemos ver que hay una gran aparente homogeneidad entre los países europeos, pues en todas las regiones europeas el número de conexiones entre países ronda el 40% del total de la región, destacando además sus conexiones con América del Sur y la zona del sudeste asiático. Esta afinidad con América del Sur se da especialmente en la parte sur de Europa, que además tiene una afinidad importante con el África del Magreb.

Si evaluamos América, vemos dos relaciones muy diferenciadas. Por un lado, Norteamérica tiene un alto nivel de afinidad con Europa, especialmente con la más industrializada Europa continental y con el sudeste asiático. Por otro, vemos cómo Latinoamérica tiene un nivel de afinidad grande con el Magreb, con la Europa mediterránea y con el sudeste asiático, zonas de similar nivel de desarrollo económico y/o socios comerciales preferentes.

Pasando ahora a Asia, podemos ver como ambas regiones en las que está dividido el continente son bastante similares, con una predominancia de los enlaces de ambas regiones entre sí y con Latinoamérica. Es interesante ver aquí la evolución de los enlaces en el contexto político-económico de cada año. Mientras que a mediados de la pasada década la paridad entre el sudeste asiático y Latinoamérica en términos de enlaces en éstas regiones

era bastante clara, en los últimos años, se viene produciendo una disminución de la afinidad entre Asia y Latinoamérica, coincidente con el parón económico de Latinoamérica durante la crisis, que no se ha visto correspondido con un parón económico similar en el Sudeste Asiático.

Si estudiamos Oceanía vemos que tiene especial relación con los países de Europa y con los del Sudeste Asiático. Podemos ver como este es un caso de clara influencia histórico-cultural (ya que estos países fueron anteriormente colonias europeas) y territorial (ya que muestra importante afinidad con la región contigua del sudeste asiático).

Por último, si nos centramos en África, vemos como hay una clara influencia territorial (la parte norte con la parte sur), posiblemente cultural-religiosa (dada la relación de ambas con oriente medio), económica (enlaces con Latinoamérica) y colonial (enlaces con Europa, especialmente la Europa continental y mediterránea).

3.5 Estudio de la entropía de los datos:

Como último experimento, se ha realizado un estudio entrópico de los datos por sector, intentando buscar una forma de ayudar a una caracterización del problema de la corrupción (determinar características que permitan determinar si un país es corrupto o no).

A continuación, se muestran los resultados de la prueba:

parlamento	justicia	policía	empresas	hacienda	aduana	medios	sanidad	educación	Servicios públicos	ejército	ONGs	religión	registro
0.75	0.76	0.71	0.82	0.85	0.85	0.91	0.89	0.91	0.82	0.94	0.98	0.99	0.96

Como se puede ver, los niveles de entropía del conjunto de datos son bastante elevados, lo cual invita a plantear posibles nuevos métodos de estudio, principalmente creando un nuevo conjunto de datos separando cada ámbito en varias características, por ejemplo, se podría separar la religión en dos campos, si el país tiene religión oficial o no y cuál es la religión predominante, el parlamento se podría diferenciar este unicameral y bicameral, la sanidad en si es publica universal, publica reducida o privada etc.

A este conjunto de datos se le aplicarían las clases expuestas en la sección 3.1.2 y se le aplicaría algún algoritmo de machine learning, por ejemplo, Naïve Bayes, para ver los valores probabilísticos asignados a cada variante, o árboles de decisión para ver que secuencia más compleja de datos caracterizaría a un país corrupto y cuál a uno que no lo es.

4 Predicción

4.1 Análisis de la metodología empleada

4.1.1 Preparación de los ficheros de predicción

Para la realización de la predicción se estudiaron diversas aproximaciones en la conformación de los ficheros de entrenamiento/test (véase sección 5.2), para, finalmente optar por emplear vectores de aprendizaje consistente en un año, un país, la región a la que pertenece dicho país, la información de los sectores y la clase asignada a dichos valores.

Se ha optado por separar los conjuntos de entrenamiento/test en función del año, quedando el año 2013 como año de test y los años 2004,2005,2006 y 2010 como años de entrenamiento.

```
2005,Argentina,AmerSur,4.60,4.50,4.30,4.30,3.60,3.40,4.20,3.40,3.00,3.00,3.30
2006,Argentina,AmerSur,4.40,4.30,3.70,4.20,4.20,3.40,3.20,2.80,2.70,3.40,3.10
2009,Argentina,AmerSur,4.86,5.02,4.91,4.75,4.92,4.86,4.85,5.09,4.75,4.99,4.86
```

Figura 32: Ejemplo de fichero de entrenamiento

Por último, se conformó el conjunto de datos para la predicción, en el cual se incluyen los mismos campos, para los años 2016-2020. Como los datos de los sectores son desconocido se marcan como tal con un '?', al igual que la clase.

```
2016,Afganistan,AsiOrM,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
2017,Afganistan,AsiOrM,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
2018,Afganistan,AsiOrM,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
2019,Afganistan,AsiOrM,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
2020,Afganistan,AsiOrM,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
```

Figura 33: Ejemplo de fichero de predicción

4.1.2 Optimización del entrenamiento

Una vez realizado el pre procesamiento de los datos de cara a la predicción, se procedió a realizar un estudio de cómo optimizar el proceso de entrenamiento, para lo cual se tienen en cuenta dos variables para un clasificador de tipo perceptrón multicapa *backpropagation*, el número de iteraciones de entrenamiento y el valor de la tasa de aprendizaje.

Este proceso de optimización se hizo trabajando con cada valor por separado, una optimización en conjunto, variando el valor de la tasa de aprendizaje para cada valor del número de iteraciones, posiblemente hubiera resultado más preciso, pero computacionalmente era mucho más costoso y se optó finalmente por ésta aproximación.

Por ello, se comenzó realizando la evaluación de la evolución de la tasa de error en el test según el número de iteraciones de entrenamiento empleadas. Se probó con una serie de iteraciones entre 100 y 1000, con saltos de 50 iteraciones entre una prueba y otra. Los resultados muestran cómo se produce una estabilización de la tasa de error a partir de las 500

iteraciones, con lo que se considera que no es necesario realizar pruebas con un número mayor del mismo.

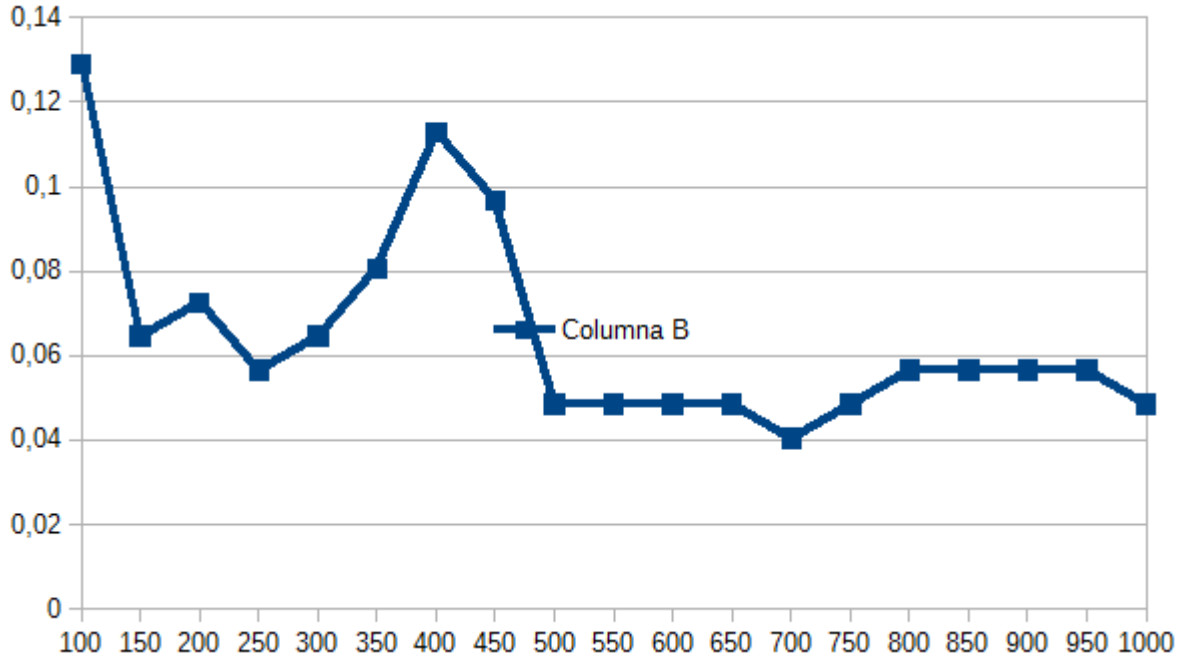

Figura 34: Evolución de la tasa de fallo

A continuación, se procedió a optimizar la tasa de aprendizaje (alfa). Para ello se fijó el número de iteraciones en 700, debido a que se trataba del valor más bajo alcanzado, y se iteró con valores de alfa entre 0.05 y 0.6 con saltos de 0.05.

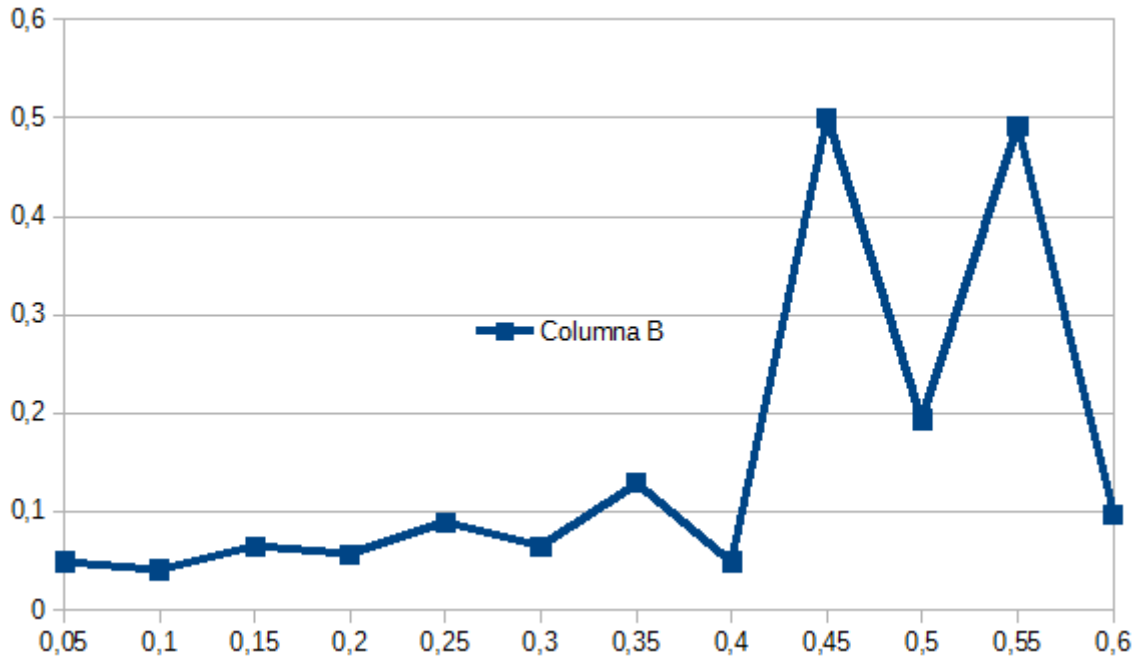

Figura 35: Evolución de la tasa de fallo

Si nos fijamos en la gráfica, podemos ver cómo los valores más bajos de la tasa de aprendizaje son los que, a su vez, producen una menor tasa de fallo, dándose el mínimo en 0.1 y produciéndose un repunte notorio a partir de 0.4.

Esto se explica porque una tasa de aprendizaje baja tiene, por lo general, una mayor de generalización que valores más grandes, con lo que el aprendizaje es más "suave" y controlado, no como puede pasar con valores más grandes, en los que el propio valor tiene una mayor volubilidad.

Tras este proceso se ha decidido realizar la predicción con un valor de tasa de aprendizaje de 0.1 y 700 iteraciones de entrenamiento.

4.2 Resultados de la predicción

4.2.1 Exposición de resultados

Si analizamos los mapas del Anexo C, podemos ver cómo se produce un decremento sustancial de la corrupción entre los años 2004 y 2013, con una tendencia a la homogeneización en la etapa de "Transition" en grandes áreas del planeta, quedando los países corruptos relativamente aislados. Resulta llamativo que hay una cierta estabilidad entre los países no corruptos, con tan solo un ligero retroceso si comparamos ambos mapas:

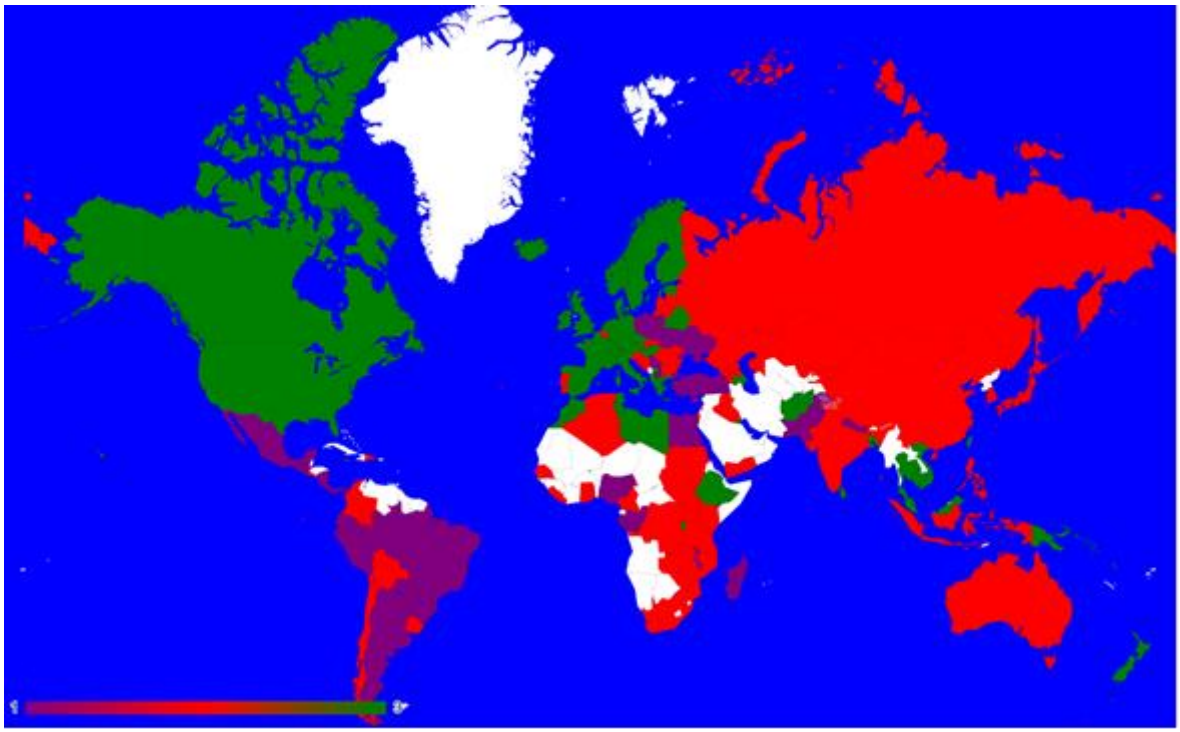

Figura 36: Clasificaciones para el año 2004

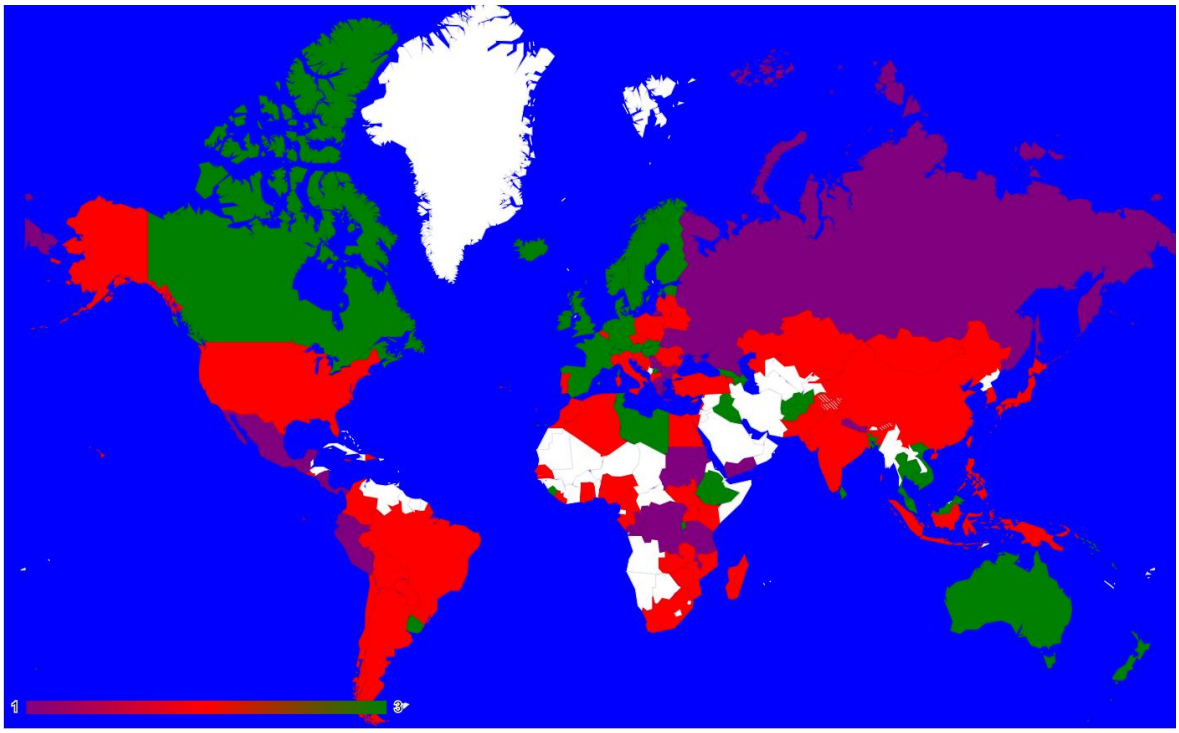

Figura 37: Clasificaciones para el año 2013

Si ahora nos fijamos en los mapas de las predicciones, podemos ver como la tendencia se mantiene a grandes rasgos y, para el año 2020, los países etiquetados como corruptos desaparecen. Podemos ver cómo, a pesar de producirse una mejoría general, ésta tiende a la homogeneización de la etapa "Transition", más que una mejoría general de los países no corruptos, que quedan prácticamente reducidos a Europa Occidental y ejemplos en la práctica aislados del resto de continentes.

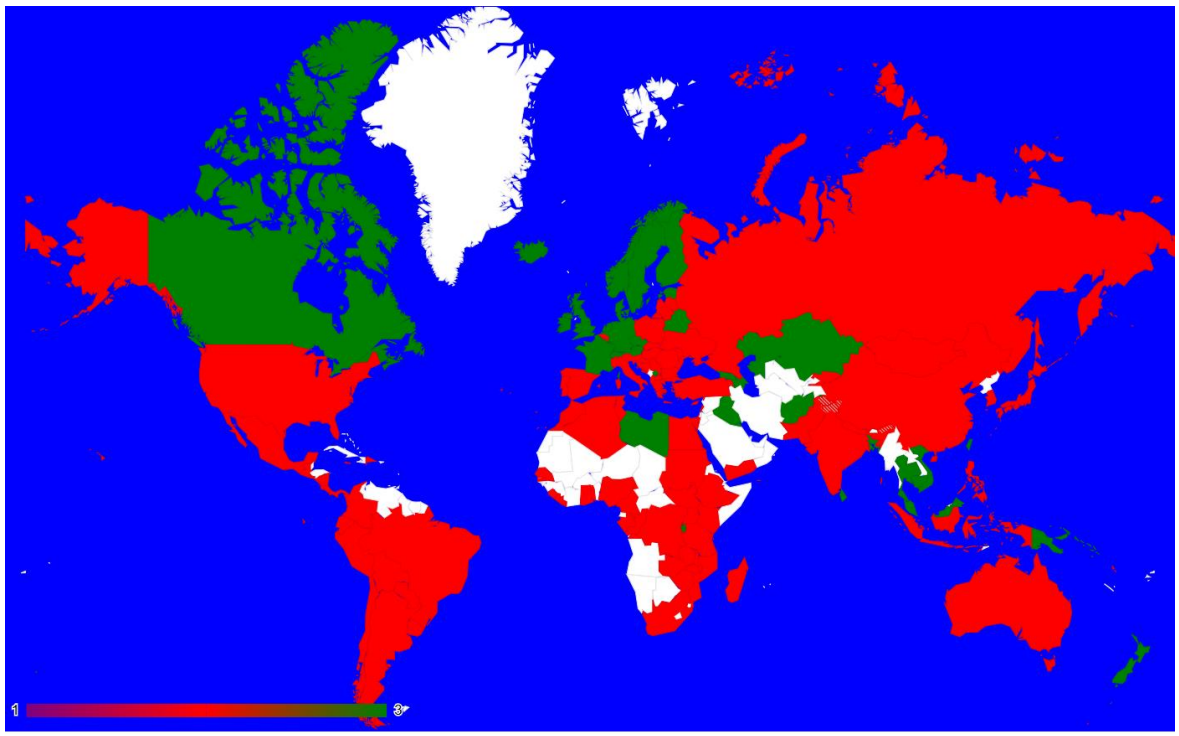

Figura 38: Clasificaciones para el año 2020

Si repasamos los resultados obtenidos en la sección 3.3 podemos encontrar una explicación a los resultados obtenidos. En esa sección se demostraba (mediante la clusterización con todos los datos a la vez) que los parámetros por los que se medía la corrupción eran variables con el tiempo.

Es por esto que, al intentar hacer una predicción en base a la tabla de datos completa, se homogeniza ese criterio, con lo que, comprensiblemente, el número de países agrupados en torno a la media aumenta, dándose un gran número de países en etapa de "Transition".

4.2.2 Exponente de Lyapunov aplicado a la predicción:

Una vez realizada la predicción es interesante estudiar el desarrollo de los valores que conforman la serie temporal final. Para esto, se emplea el exponente de Lyapunov.

A la hora de calcular el exponente de Lyapunov para el conjunto se ha empleado como medición de la distancia la diferencia entre las clases de cada vector, no pudiéndose realizar el cálculo sobre los valores del BGC debido a ausencia de tales para los años 2016-2020. De este modo, se otorgó a la clase *Corrupt*, el valor 0, a la clase *Transition* el valor 1 y a la clase *Clean*, el valor 2. A continuación se muestran los resultados:

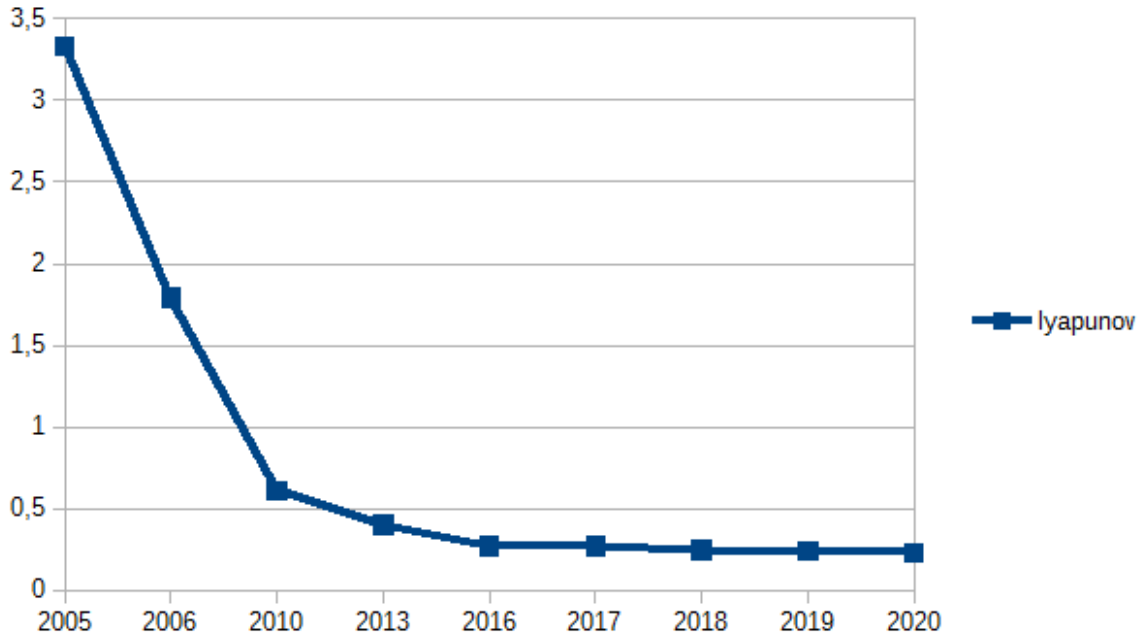

Figura 39: Evolución del exponente de Lyapunov

Como podemos ver, los valores obtenidos son siempre mayores que 0 y, por tanto, según lo expuesto en la sección 1.4.2, caóticos, aunque hay que destacar que la tendencia de dichos valores es descendente, por lo que, como quedaba expuesto en esa misma sección, existe una tendencia al orden. De aquí se puede extraer que, en la medida en que los acontecimientos son predecibles, la tendencia será asintótica, acercándose al orden (Lyapunov 0) cada vez más.

Como complemento a esta evaluación se ha realizado el mismo cálculo para los años 2005,2006,2010 y 2013 usando las diferencias entre los valores del BGC para obtener un

resultado más preciso, que se espera sirva para clarificar el grado de aproximación a la realidad de los valores antes mostrados. A continuación se muestran los resultados:

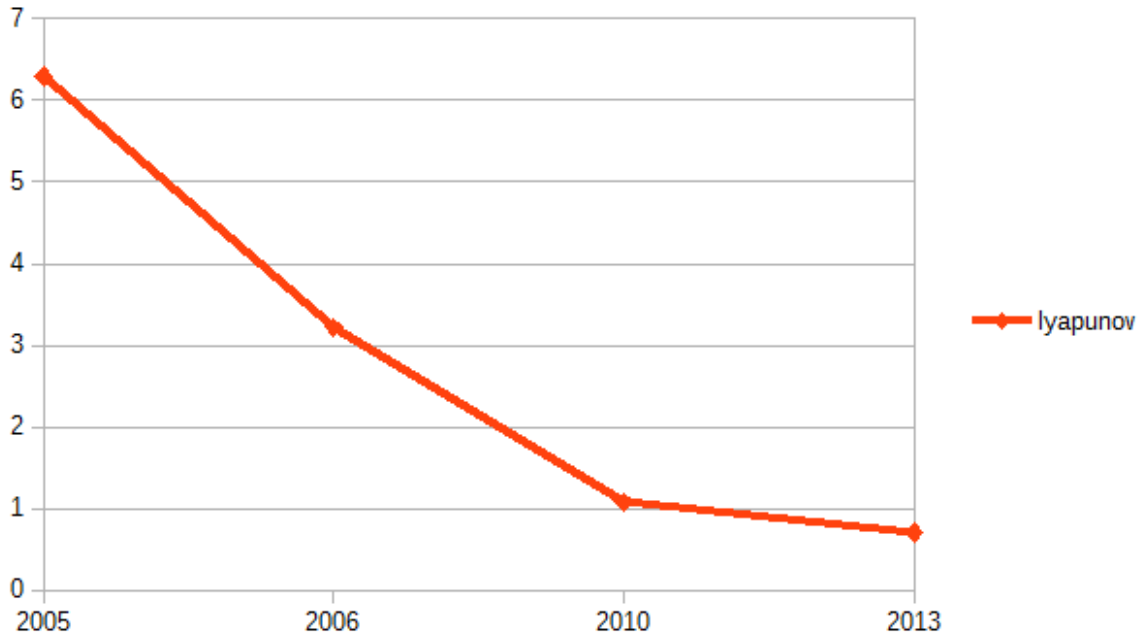

Figura 40: Evolución del exponente de Lyapunov

Como podemos ver, el comportamiento del exponente de Lyapunov en este caso es muy similar al del caso anterior, De hecho, si vemos, el grado de diferencia entre los valores de un año y otro es relativamente constante:

Lyapunov Valores	Lyapunov Clases	Factor de diferencia
6,29	*3,33*	*1,88888888888889*
3,22	*1,79*	*1,79888268156425*
1,08	*0,61*	*1,77049180327869*
0,71	*0,4*	*1,775*

Tabla 2: Factor de diferenciación

Estos resultados sirven para reforzar los obtenidos previamente, otorgando así mayor validez a las conclusiones extraídas de los mismos.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

El estudio de la corrupción es un campo de especial interés sociológico. A lo largo del presente Trabajo de Fin de Grado se ha pretendido diseñar una metodología para el estudio y predicción de su evolución, así como un inicio del proceso de caracterización de la corrupción en el mundo.

Para ello se he comenzado por un estudio de los datos, consistente en aplicar el algoritmo *backpropagation* sobre distintas formas de presentar los datos (y con objetivos de predicción distintos a su vez), obteniendo como principal resultado el conocimiento de que hay, al menos en el conjunto de datos de que se disponía, una diferenciación reconocible entre los países de una misma región. Como siguiente paso se ha procedido a realizar dos clusterizaciones, sobre 2 y 3 *clusters*, respectivamente, con objeto de inferir relaciones entre países asignados a los mismos grupos. De éste proceso se obtuvieron las conclusiones de que hay, a grandes rasgos, una relación entre la posición geográfica y el nivel de corrupción, se estableció la posibilidad de que ello tuviera que ver con la influencia de las grandes potencias regionales, y se comprobó que se había producido una mejoría sustancial en los niveles de corrupción mundiales en el período 2004-2013. Finalmente se hicieron dos pruebas más, el establecimiento de enlaces entre países que permitió establecer hipótesis sobre los distintos medios de influencia (comercial, colonial, cultural…) que podían afectar al grado de corrupción de un país, y el cálculo de las entropías de los distintos sectores que sirvió para establecer posibles rutas de cara a continuar la investigación aquí comenzada.

Ya en la segunda mitad del trabajo, se procedió a realizar un proceso de optimización del entrenamiento de la red neuronal, consistente en dos fases: una primera en la que se optimizaba el número de iteraciones de entrenamiento de la red, para posteriormente optimizar el valor del factor de aprendizaje. Posteriormente se realizó la clasificación para los años del período 2016-2020, dando como resultado una evolución temporal que seguía mostrando una mejoría general, con la práctica erradicación de países clasificados como *Corrupt* para el año 2020. Finalmente se procedió a realizar dos cálculos de la evolución de la serie temporal, el primero tomando sólo las diferencias entre clases, tanto del conjunto original como de los años predichos, y el segundo tomando las diferencias de los valores de cada sector, utilizando sólo los valores de los años originales para estudiar el grado de aproximación de los resultados usando sólo clases con éstos, dando como resultado una desviación medianamente estable que permitía prever que la evolución de la predicción seguiría la misma tendencia que se había obtenido.

5.2 Trabajo futuro

A lo largo del desarrollo del presente trabajo se han planteado múltiples opciones, bien de estudio de los datos, bien de modelos de predicción, que no han podido ser desarrollados y que, por tanto, se dejan como trabajo futuro para futuras ampliaciones del mismo.

Una primera opción, derivada del estudio de los datos realizado en la sección 3.2, sería la de realizar una clusterización usando tantos clústers como regiones hay, viendo así si se puede reafirmar la conclusión expuesta en dicho punto acerca de la relación, en términos generales, entre la posición geográfica y un determinado nivel de corrupción.

Otra de las opciones planteadas en el trabajo, detallada en parte en la sección 3.5 es la conformación de una nueva base de datos , compuesta por datos no directamente relacionados con la corrupción pero sí suficientes como para caracterizar un Estado, como puedan ser el sistema de gobierno, la existencia o no de sanidad pública universal o el nivel impositivo medio, a los que se asignarían los valores de corrupción del conjunto de datos actual para así intentar extraer una imagen de ciertas condiciones que, al menos estadísticamente, llevan a un país a ser considerado corrupto, además de permitir ampliar con datos de años que no se encontraban en el conjunto original, pues los valores sí que podrían conocerse para esta base de datos, permitiendo así, quizá, una predicción más completa y extensa que la realizada e incluso el estudio de pioneros y seguidores en la serie temporal.

De manera alternativa, se plantean también nuevas posibles formas de abordar la parte de la realización de una predicción sobre el conjunto de datos. La primera de ellas es probar con otros métodos de aprendizaje automático, como puedan ser la regresión logística o el algoritmo de vecinos próximos que, se considera, podrían dar buenos resultados teniendo en cuenta diferentes características de los datos actuales, como puede ser, por ejemplo, la necesaria clasificación de dos vectores suficientemente similares con la misma clase. También se plantea el empleo de distintos tipos de red neuronal, como *MADALINE*.

Otra alternativa a la hora de predecir valores de corrupción futuros es ir realizando la predicción campo a campo. Esto requeriría pasar por más procedimientos de entrenamiento/test -uno por campo- y presenta posibles problemas, como que se ignoraría una posible interdependencia entre campos, pero podría aportar resultados interesantes y más completos que los ya obtenidos.

Adicionalmente, podría plantearse la predicción de los datos en el contexto de una serie de temporal, aplicándose un proceso de predicción basado en éstas para cada uno de los campos.

Como última manera alternativa de abordar la predicción se plantea el uso de una red bayesiana. Esto presenta el problema de que la probabilidad de que un valor de un año que no se ha incluido sea X es 0. Ello se podría arreglar mediante el corrector de *Laplace* o, y posiblemente ésta opción pueda resultar más interesante, aproximando el valor de dichas probabilidades a una función matemática dependiente del año, suponiendo que las tendencias ocurridas en el período de tiempo que contemplan los datos se mantienen.

Además, se plantean, como ya se comentaba en la sección 4.1.2, realizar la optimización de aprendizaje empleando los dos parámetros a la vez, pero reduciendo el número de iteraciones en ambos casos, por ejemplo, se podría probar los valores de tasa de aprendizaje entre 0.05 y 0.25 y con valores entre 400 y 900 para las iteraciones.

La última alternativa planteada, a partir de los resultados de la sección 4.2 es la de establecer una ligera variación en el mecanismo de asignación de clase de la sección 3.1.2, realizando la asignación en función de la media y la varianza de los vectores que pertenecen al mismo año que el país que se quiere clasificar, en lugar de los valores del conjunto al completo, para después aplicar la metodología de optimización descrita en la sección 4.1.2 o la antes descrita en ésta misma sección y realizar la predicción nuevamente con la red neuronal o con cualquiera de los otros métodos presentados aquí.

Glosario

API	Application Programming Interface
ADALINE	Adaptative Linear Network
MADALINE	Many Adaptative Linear Network
CIS	Centro de Investigaciones Sociológicas
GNU	GNU's Not Unix

Anexos

Anexo A: Mapas de Clusterización

Año 2004

Año a Año:

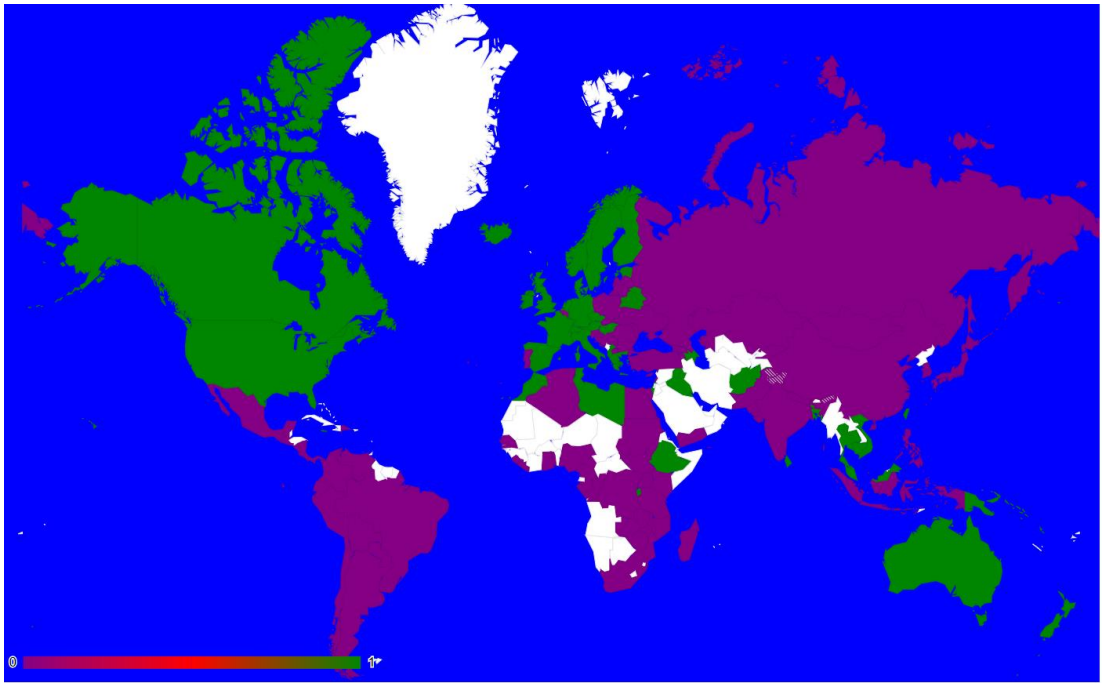

Figura 41: Agrupamiento en 2 clusters 2004

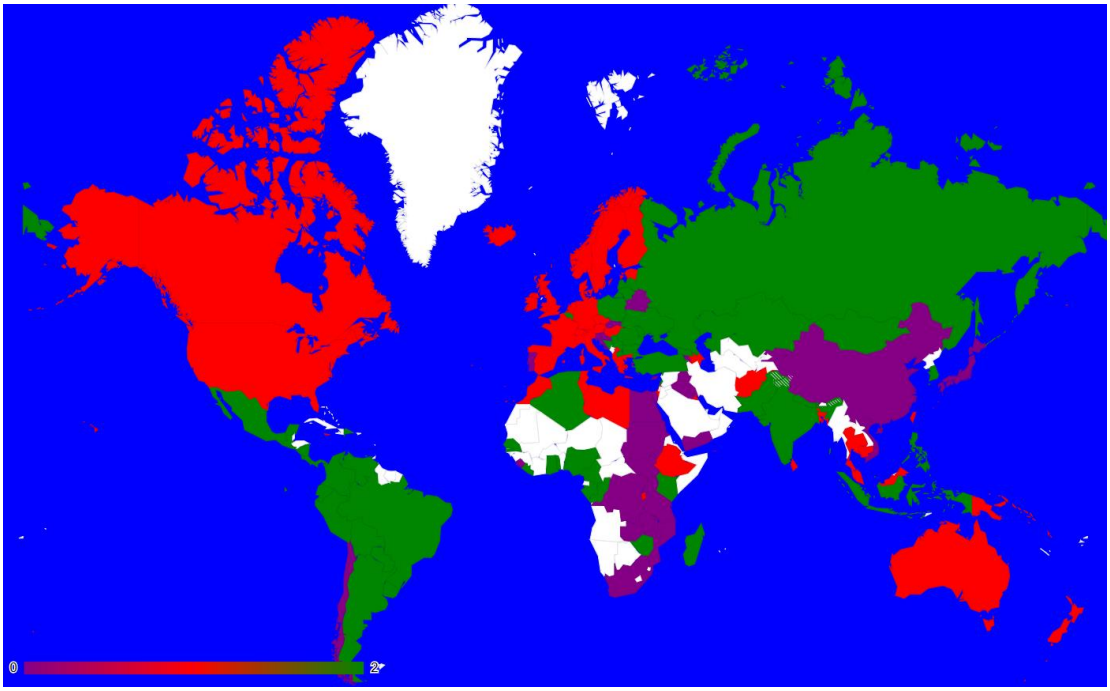

Figura 42: Agrupamiento en 3 clusters 2004

Todos los años:

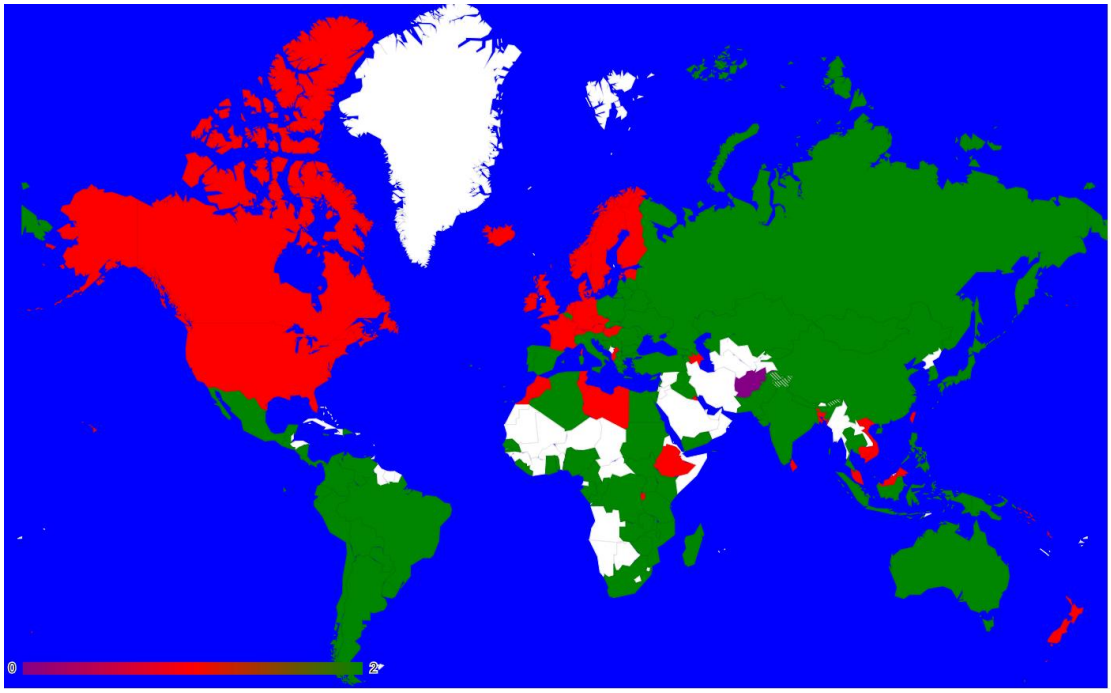

Figura 43: Agrupamiento en 2 clusters 2004

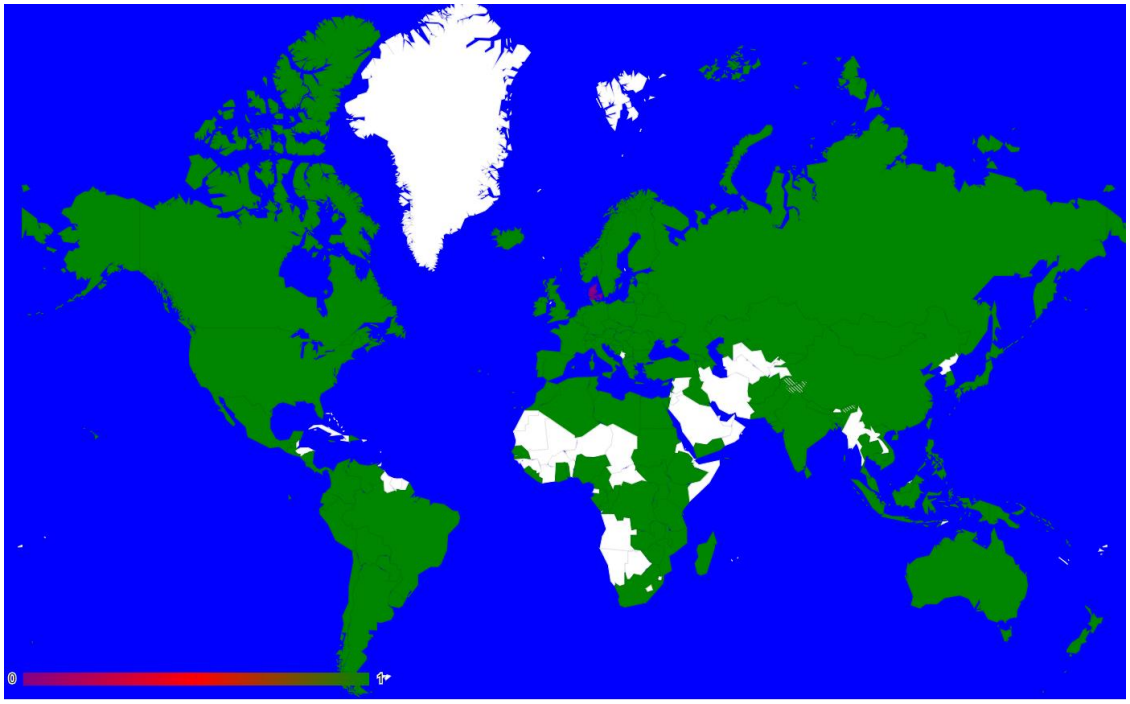

Figura 44: Agrupamiento en 3 clusters 2004

Año 2005

Año a Año:

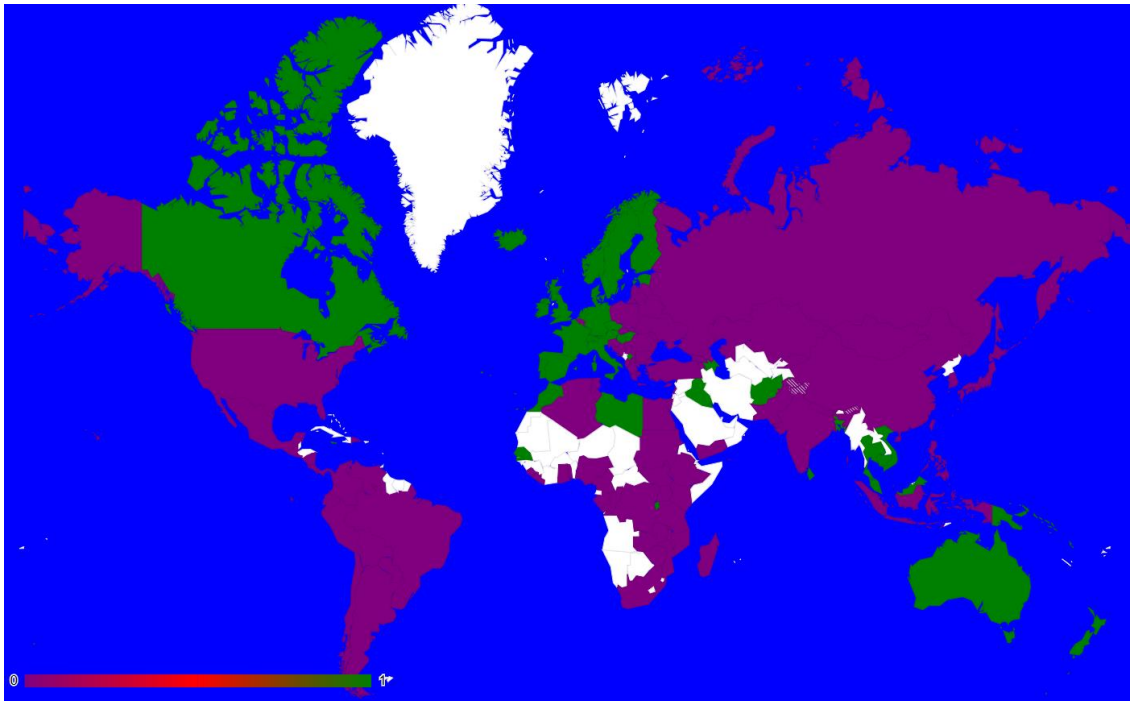

Figura 45: Agrupamiento en 2 clusters 2005

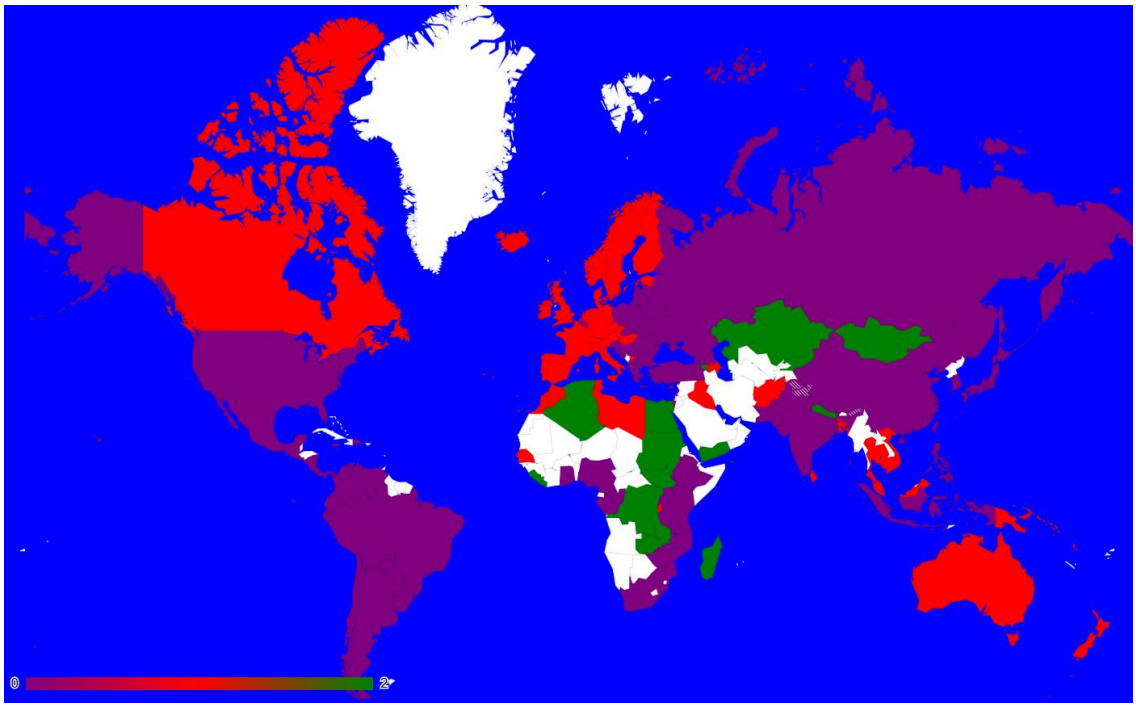

Figura 46: Agrupamiento en 3 clusters 2005

Todos los años:

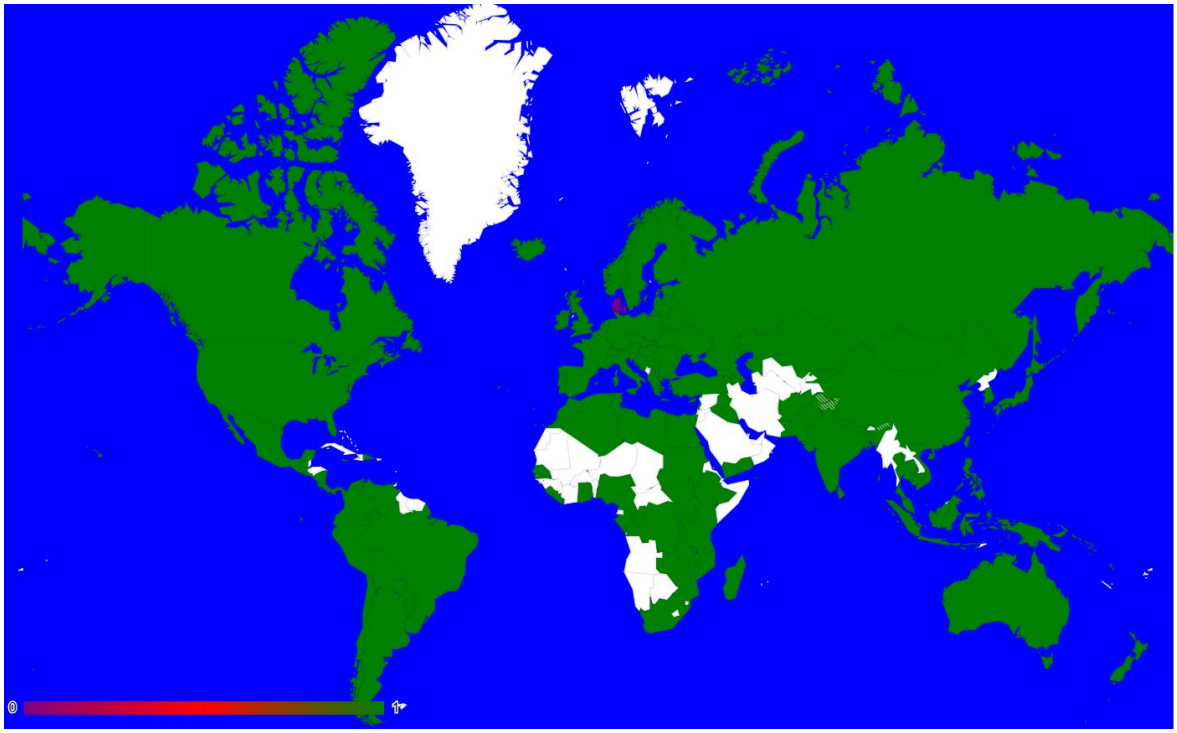

Figura 47: Agrupamiento en 2 clusters 2005

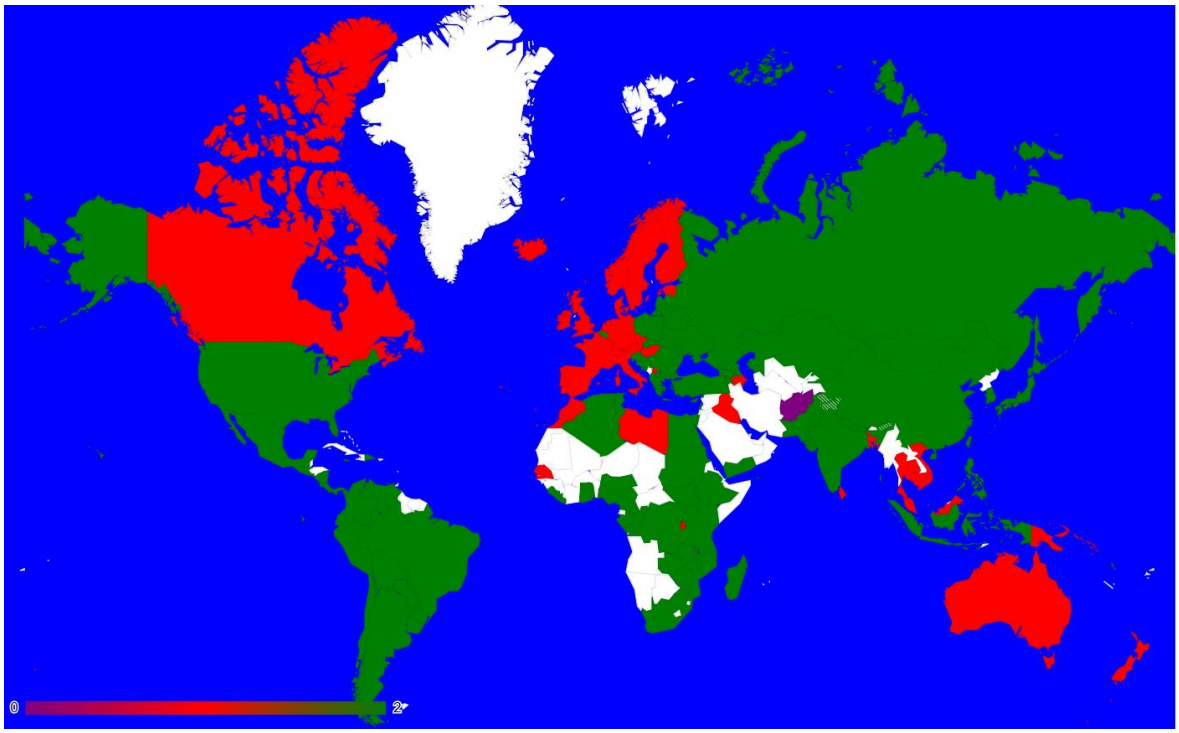

Figura 48: Agrupamiento en 3 clusters 2005

Año 2006

Año a Año:

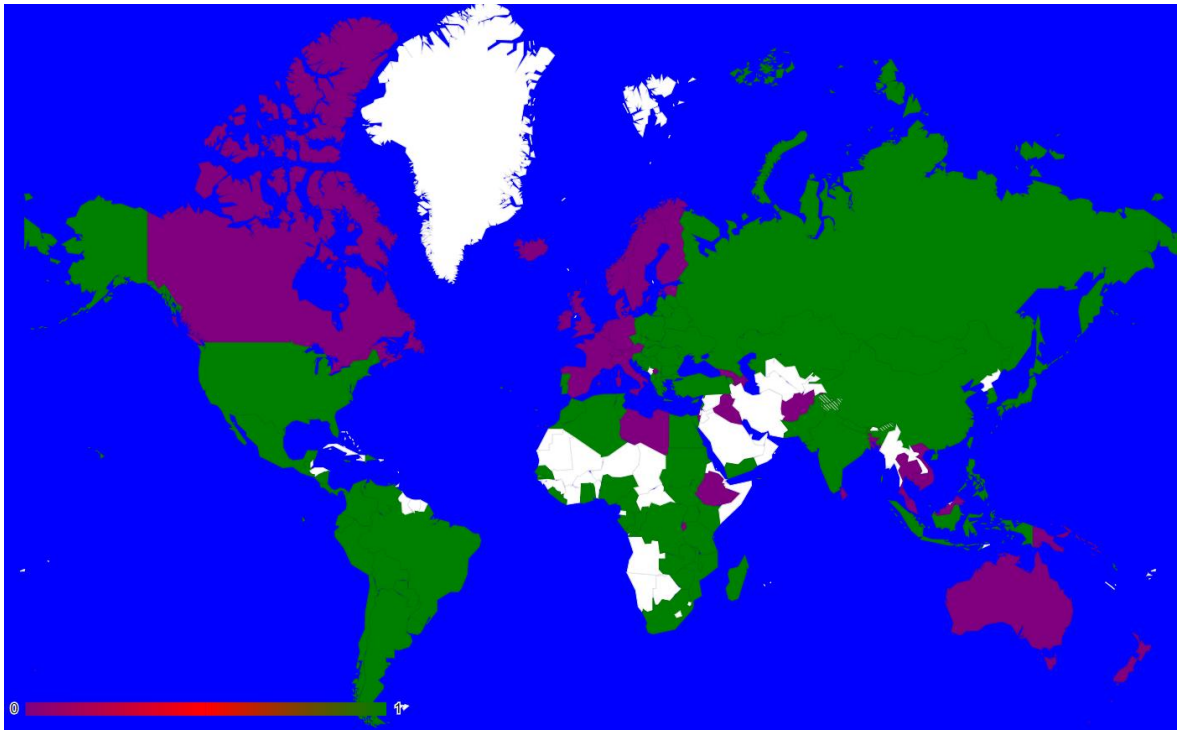

Figura 49: Agrupamiento en 2 clusters 2006

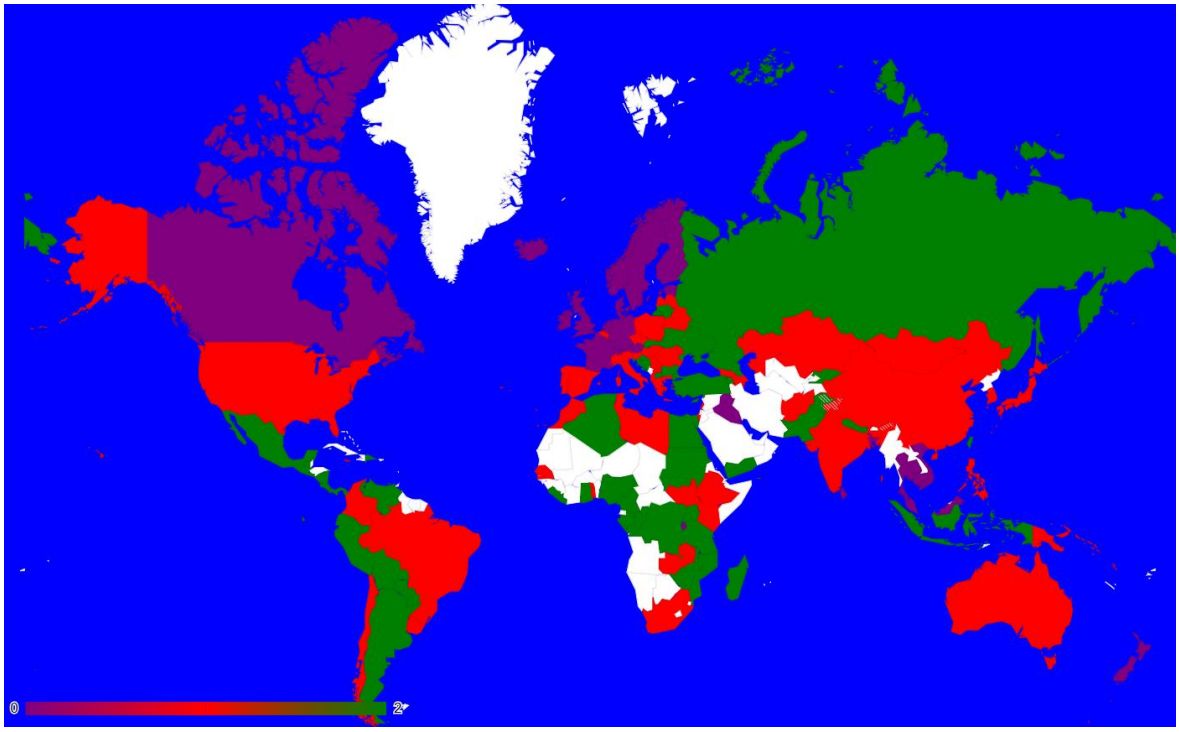

Figura 50: Agrupamiento en 3 clusters 2006

Todos los años:

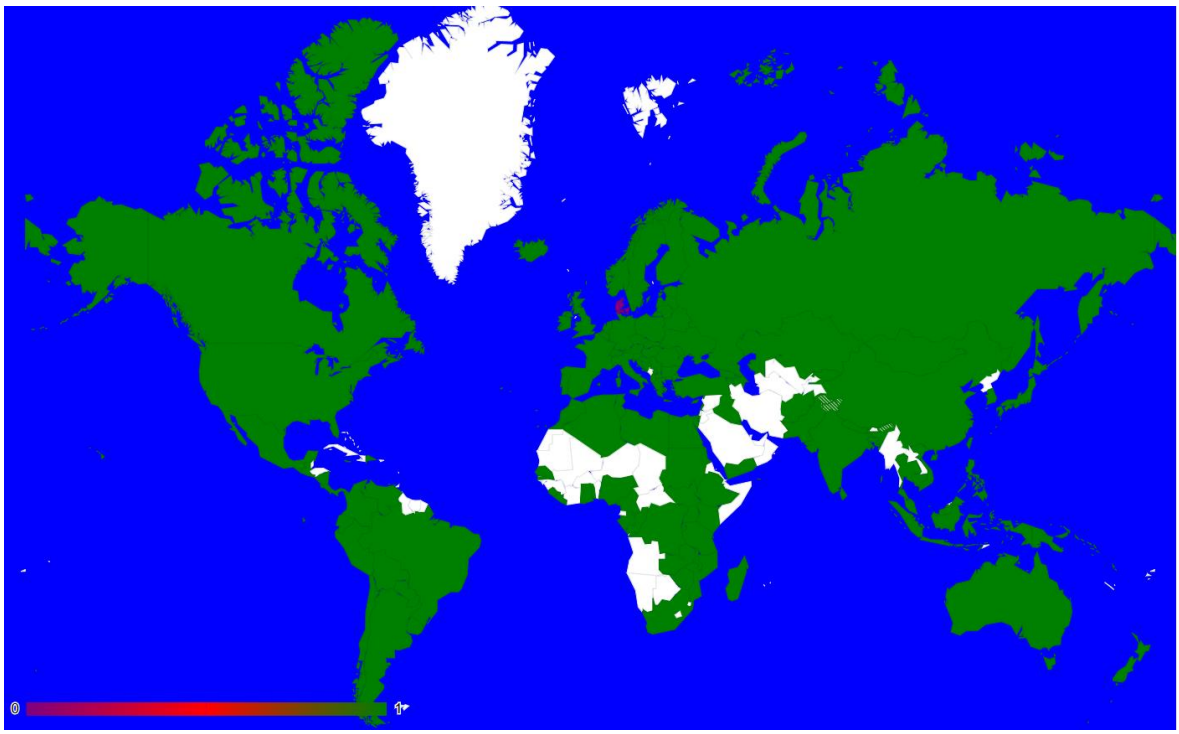

Figura 51: Agrupamiento en 2 clusters 2006

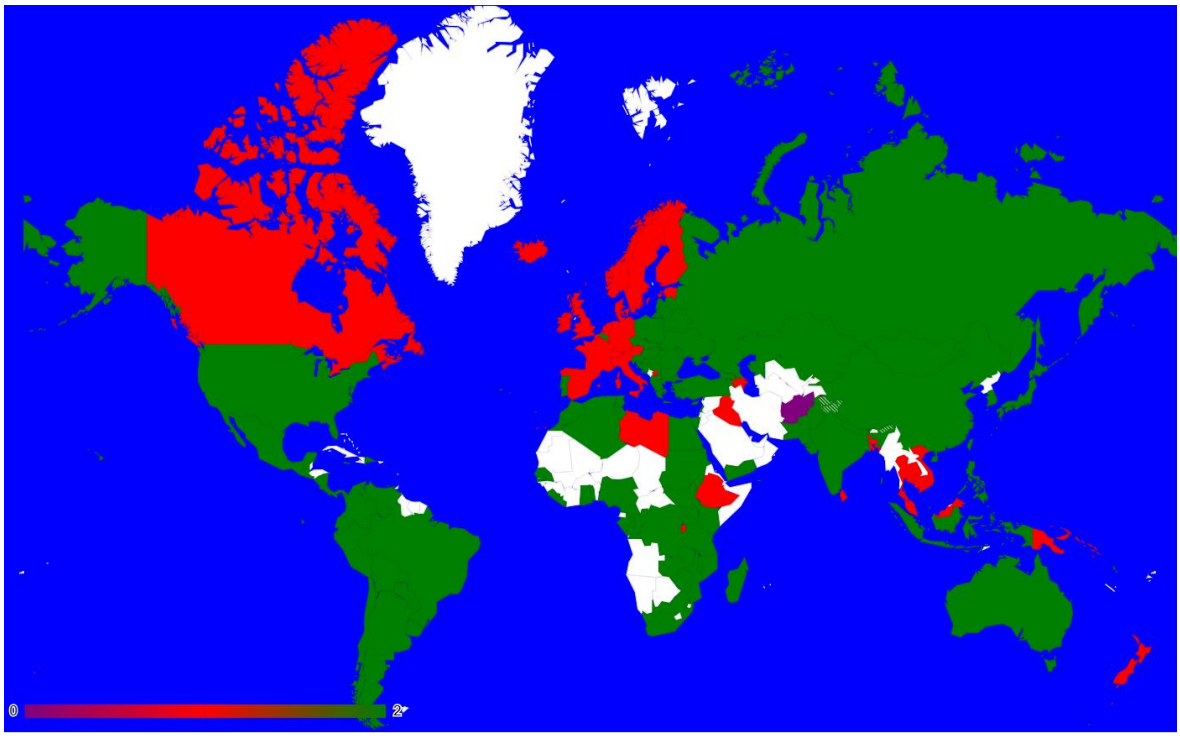

Figura 52: Agrupamiento en 3 clusters 2006

Año 2010

Año a Año:

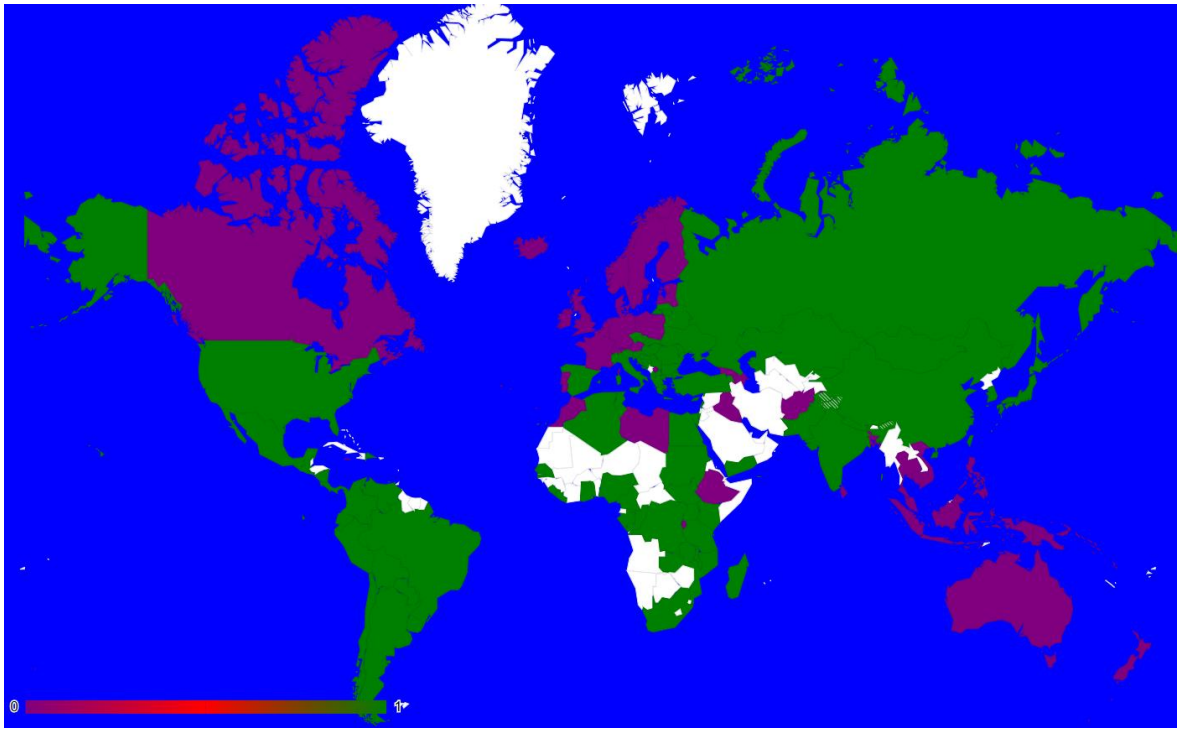

Figura 53: Agrupamiento en 2 clusters 2010

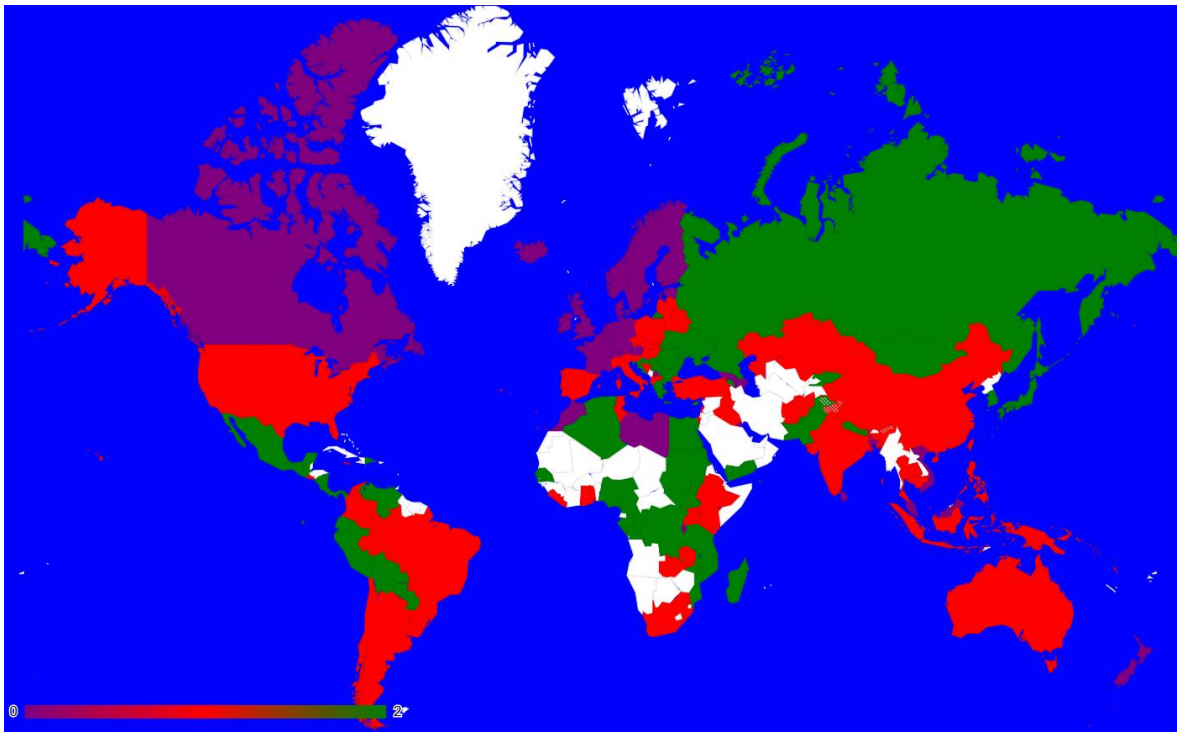

Figura 54: Agrupamiento en 3 clusters 2010

Todos los años:

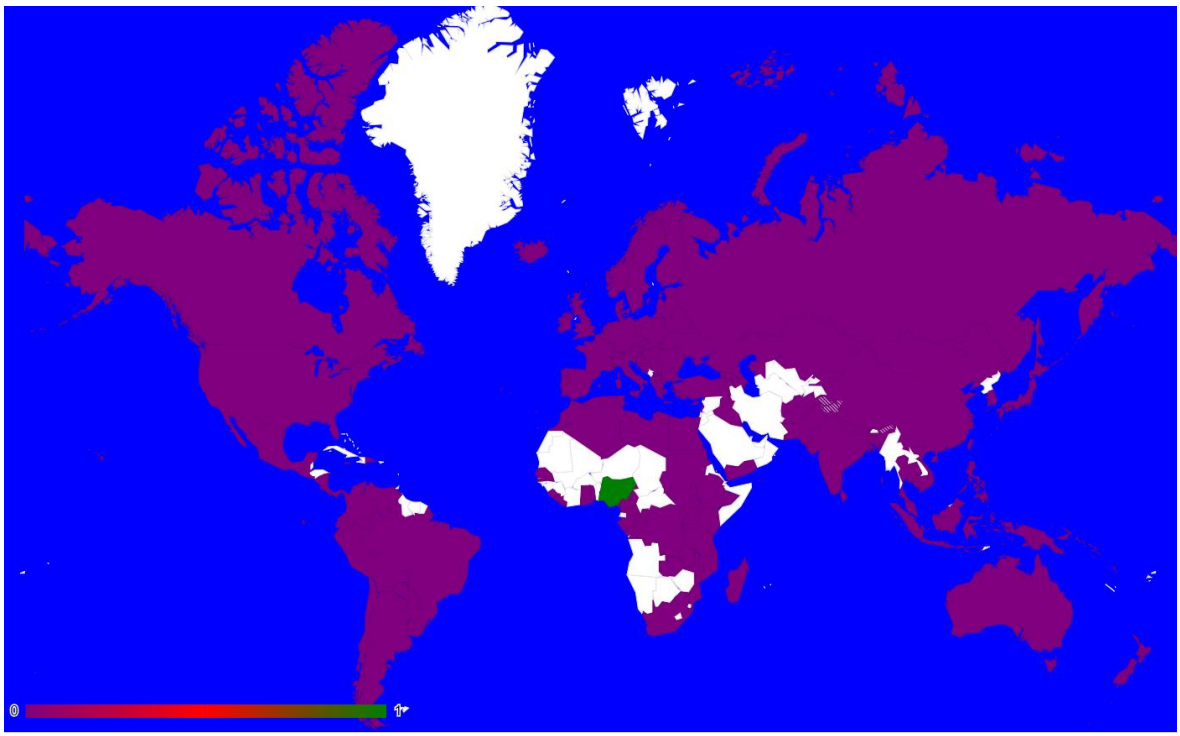

Figura 55: Agrupamiento en 2 clusters 2010

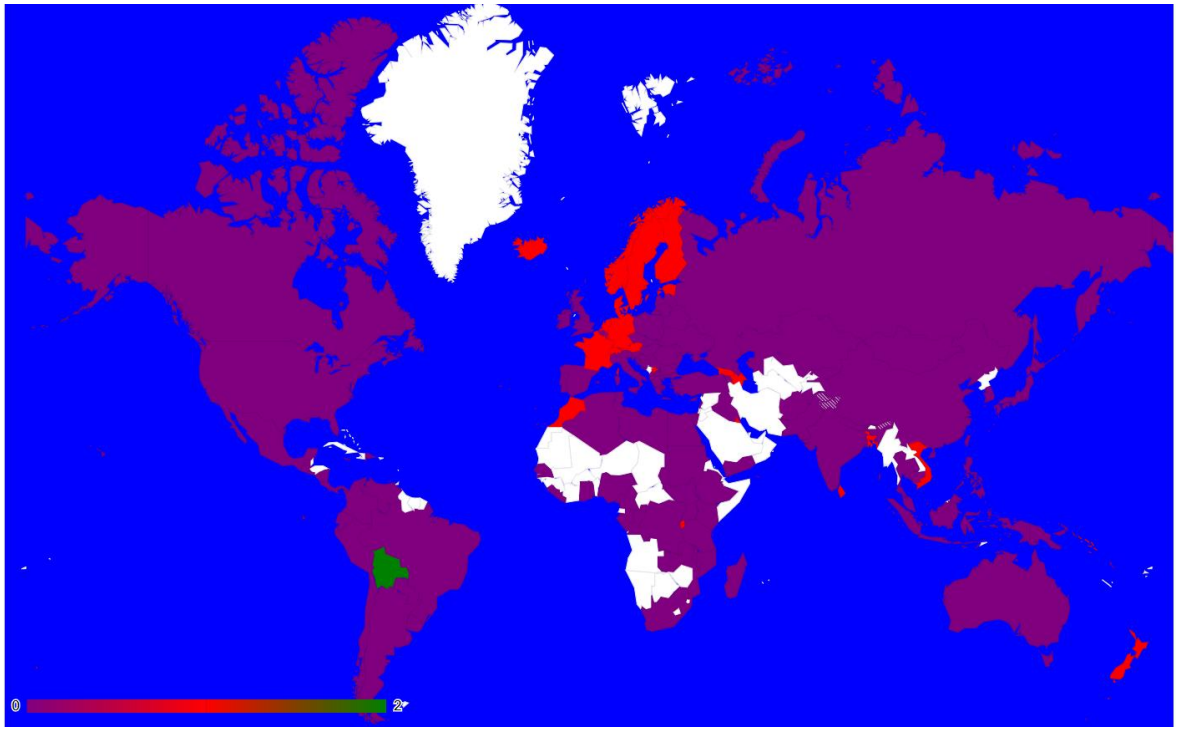

Figura 56: Agrupamiento en 3 clusters 2010

Año 2013

Año a Año:

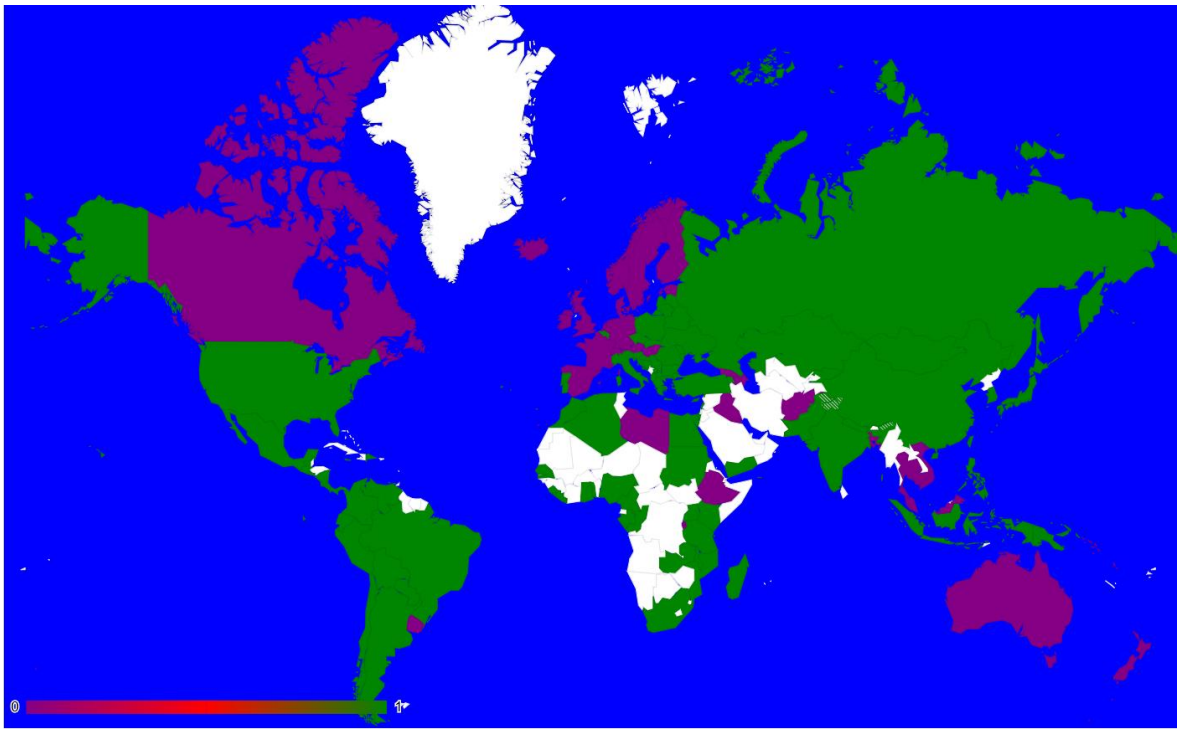

Figura 57: Agrupamiento en 2 clusters 2013

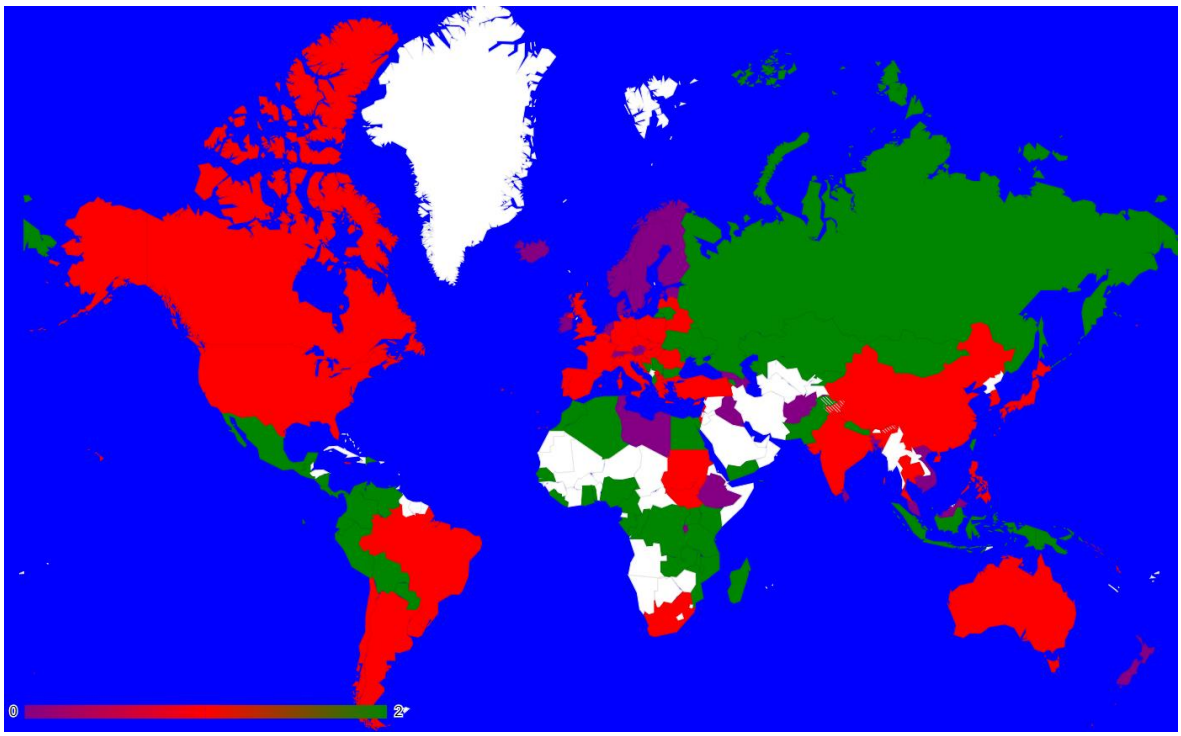

Figura 58: Agrupamiento en 3 clusters 2013

Todos los años:

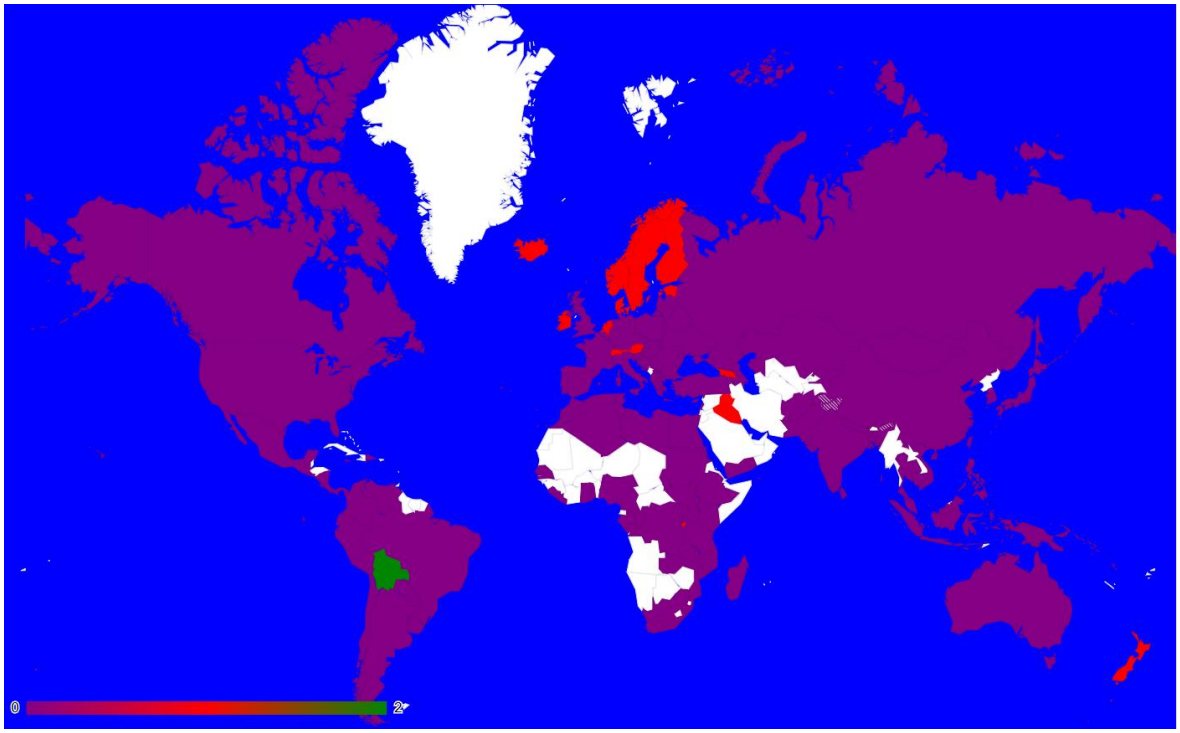

Figura 59: Agrupamiento en 2 clusters 2013

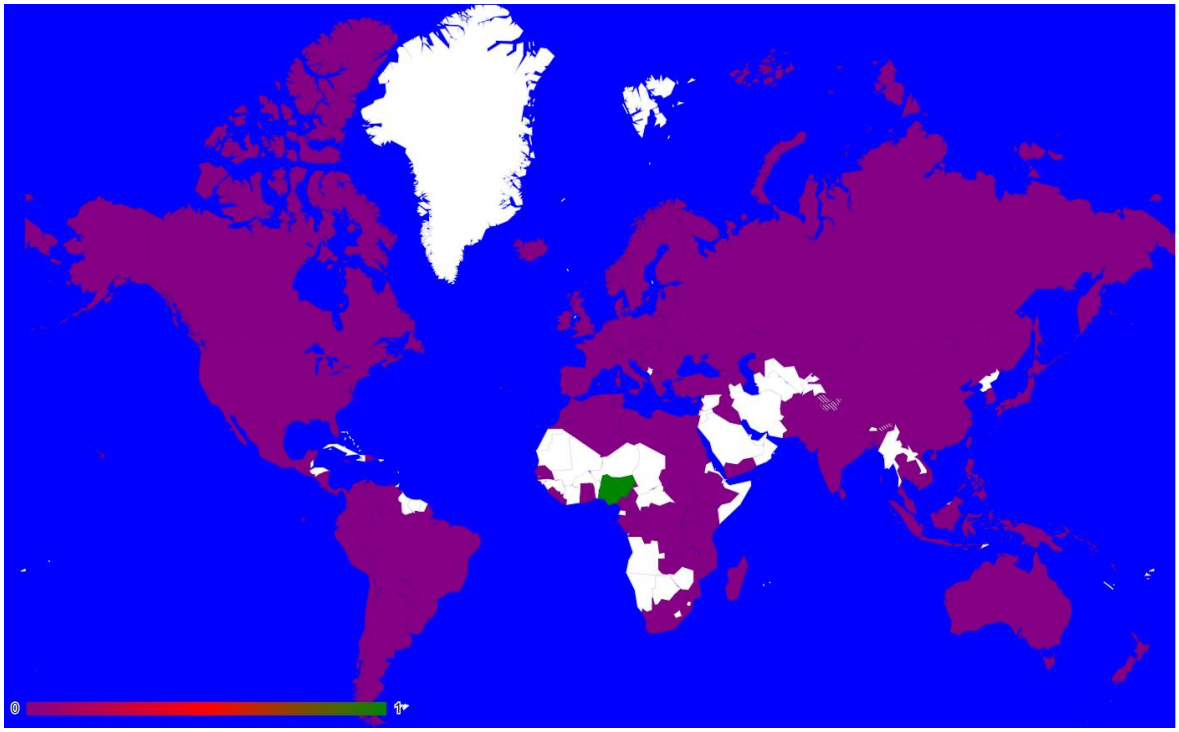

Figura 60: Agrupamiento en 3 clusters 2013

Anexo B: Enlaces

Año 2004

region	*EurCent*	*EurNor*	*EurSur*	*AmerSur*	*AmerNor*	*AsiOrM*	*IndoChina*	*Oceania*	*AfrSubSah*	*AfrSah*
EurCent	*76*	*57*	*55*	*55*	*17*	*34*	*56*	*10*	*33*	*34*
EurNor	*57*	*26*	*29*	*20*	*6*	*14*	*30*	*3*	*10*	*13*
EurSur	*55*	*29*	*50*	*66*	*10*	*30*	*50*	*3*	*32*	*38*
AmerSur	*55*	*20*	*66*	*140*	*3*	*40*	*72*	*0*	*46*	*59*
AmerNor	*17*	*6*	*10*	*3*	*2*	*5*	*10*	*4*	*3*	*1*
AsiOrM	*34*	*14*	*30*	*40*	*5*	*28*	*30*	*3*	*33*	*24*
IndoChina	*56*	*30*	*50*	*72*	*10*	*30*	*42*	*7*	*30*	*32*
Oceania	*10*	*3*	*3*	*0*	*4*	*3*	*7*	*2*	*0*	*0*
AfrSubSah	*33*	*10*	*32*	*46*	*3*	*33*	*30*	*0*	*48*	*43*
AfrSah	*34*	*13*	*38*	*59*	*1*	*24*	*32*	*0*	*43*	*24*

Tabla 3: Enlaces del año 2004

Año 2005

region	*EurCent*	*EurNor*	*EurSur*	*AmerSur*	*AmerNor*	*AsiOrM*	*IndoChina*	*Oceania*	*AfrSubSah*	*AfrSah*
EurCent	90	52	70	56	23	46	76	12	36	36
EurNor	52	46	22	22	6	14	31	7	11	20
EurSur	70	22	50	55	11	33	54	4	33	25
AmerSur	56	22	55	152	6	41	73	0	53	56
AmerNor	23	6	11	6	2	9	13	3	6	3
AsiOrM	46	14	33	41	9	36	43	4	34	25

IndoChina	76	31	54	73	13	43	62	7	29	33
Oceania	12	7	4	0	3	4	7	2	0	2
AfrSubSah	36	11	33	53	6	34	29	0	48	40
AfrSah	36	20	25	56	3	25	33	2	40	24

Tabla 4: Enlaces del año 2005

Año 2006

region	*EurCent*	*EurNor*	*EurSur*	*AmerSur*	*AmerNor*	*AsiOrM*	*IndoChina*	*Oceania*	*AfrSubSah*	*AfrSah*
EurCent	76	61	57	66	14	29	67	4	27	27
EurNor	61	32	24	27	9	14	26	3	15	15
EurSur	57	24	42	67	13	25	36	5	28	27
AmerSur	66	27	67	124	11	36	54	0	42	49
AmerNor	14	9	13	11	0	7	8	1	4	2
AsiOrM	29	14	25	36	7	18	26	5	26	17
IndoChina	67	26	36	54	8	26	44	6	29	20
Oceania	4	3	5	0	1	5	6	2	3	1
AfrSubSah	27	15	28	42	4	26	29	3	26	43
AfrSah	27	15	27	49	2	17	20	1	43	24

Tabla 5: Enlaces del año 2006

Año 2010

region	*EurCent*	*EurNor*	*EurSur*	*AmerSur*	*AmerNor*	*AsiOrM*	*IndoChina*	*Oceania*	*AfrSubSah*	*AfrSah*
EurCent	98	64	70	79	14	49	86	14	33	42
EurNor	64	34	30	23	10	22	48	4	9	12

EurSur	70	30	68	81	12	36	70	4	28	34
AmerSur	79	23	81	138	10	50	102	2	45	65
AmerNor	14	10	12	10	2	8	20	2	2	1
AsiOrM	49	22	36	50	8	16	49	6	23	25
IndoChina	86	48	70	102	20	49	98	11	43	42
Oceania	14	4	4	2	2	6	11	2	5	1
AfrSubSah	33	9	28	45	2	23	43	5	38	42
AfrSah	42	12	34	65	1	25	42	1	42	28

Tabla 6: Enlaces del año 2010

Año 2013

region	*EurCent*	*EurNor*	*EurSur*	*AmerSur*	*AmerNor*	*AsiOrM*	*IndoChina*	*Oceania*	*AfrSubSah*	*AfrSah*
EurCent	104	51	73	72	24	54	73	10	34	45
EurNor	51	40	18	24	6	17	21	7	10	10
EurSur	73	18	70	88	14	42	53	0	30	37
AmerSur	72	24	88	172	9	55	86	0	52	72
AmerNor	24	6	14	9	2	11	12	2	2	4
AsiOrM	54	17	42	55	11	30	50	2	26	27
IndoChina	73	21	53	86	12	50	62	7	34	40
Oceania	10	7	0	0	2	2	7	2	3	3
AfrSubSah	34	10	30	52	2	26	34	3	32	40
AfrSah	45	10	37	72	4	27	40	3	40	50

Tabla 7: Enlaces del año 2013

Anexo C: Mapas de Clasificación

Año 2004

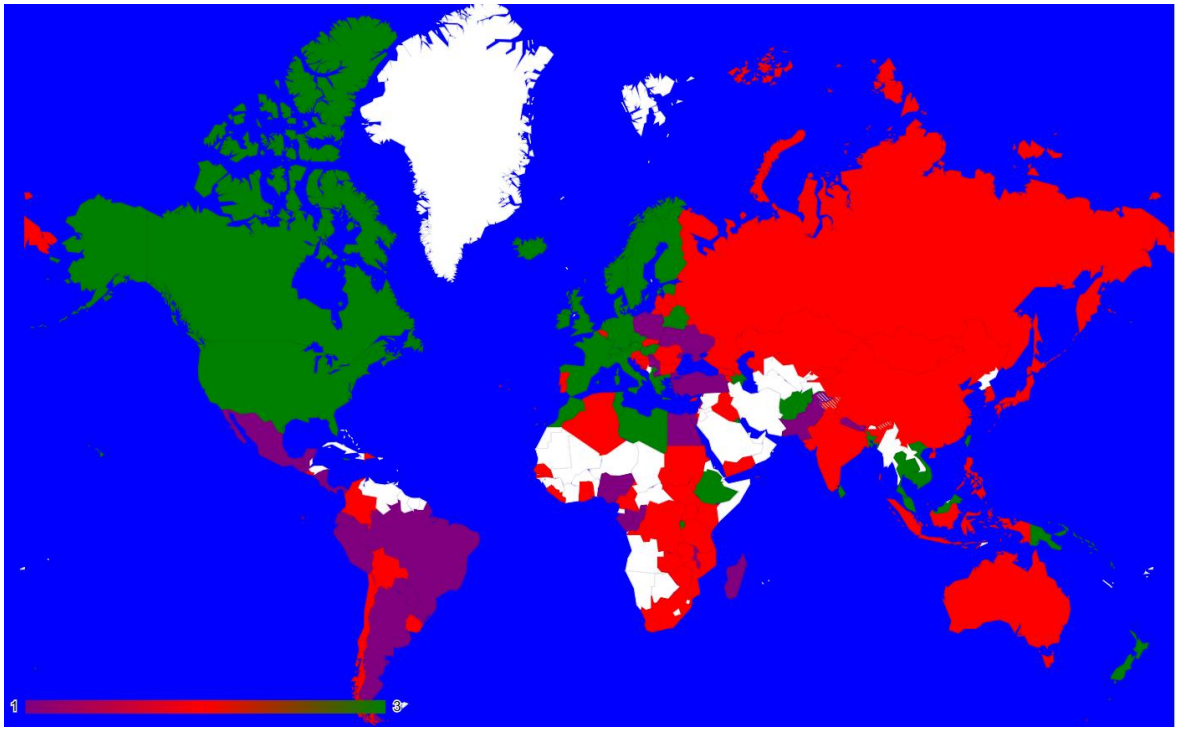

Figura 61: Clasificación año 2004

Año 2005

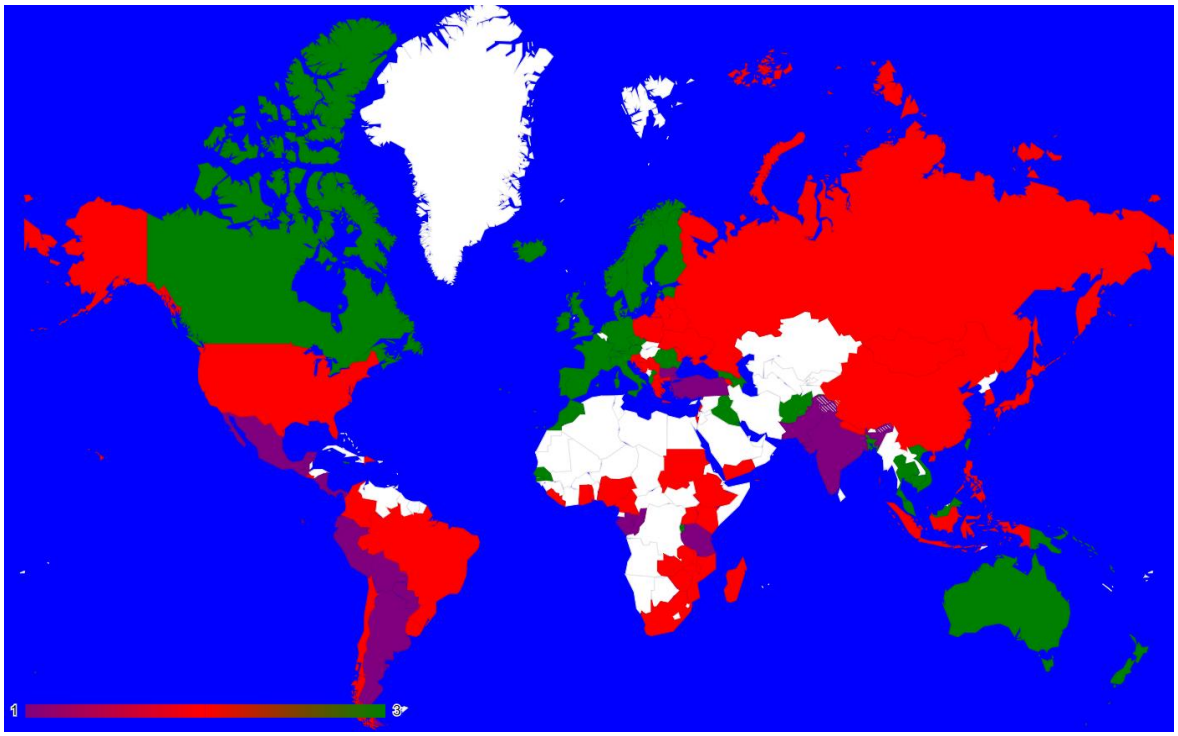

Figura 62:Clasificación año 2005

Año 2006

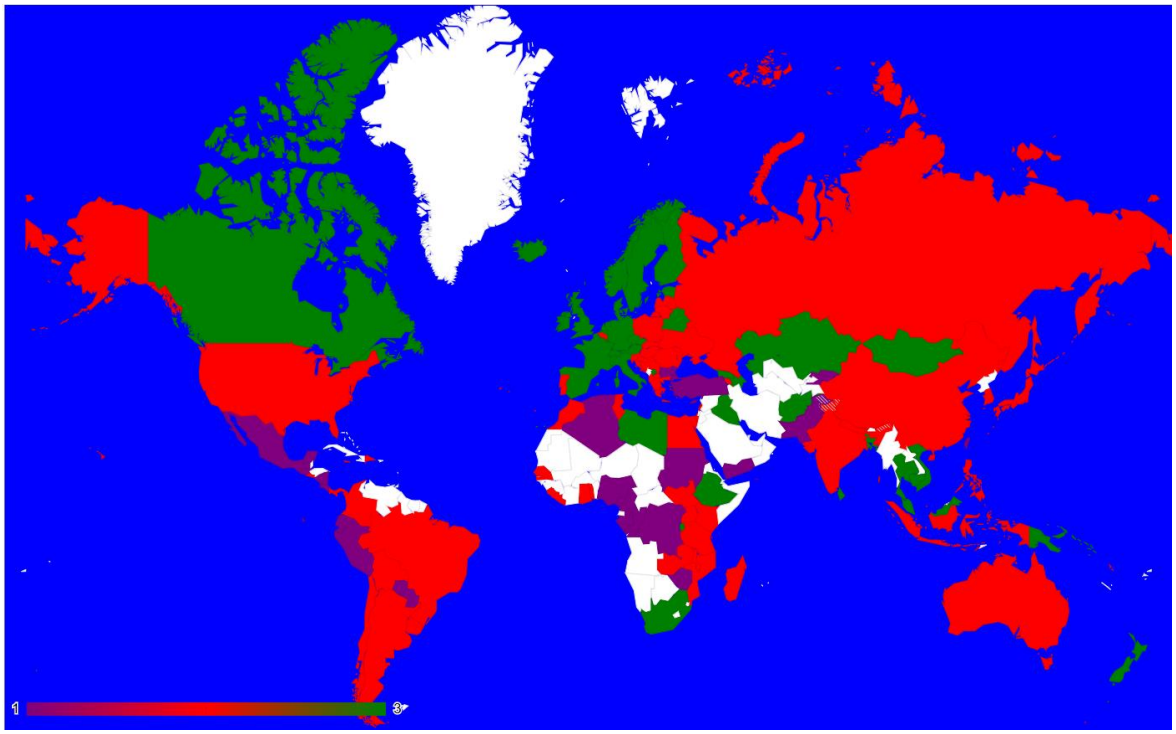

Figura 63: Clasificación año 2006

Año 2010

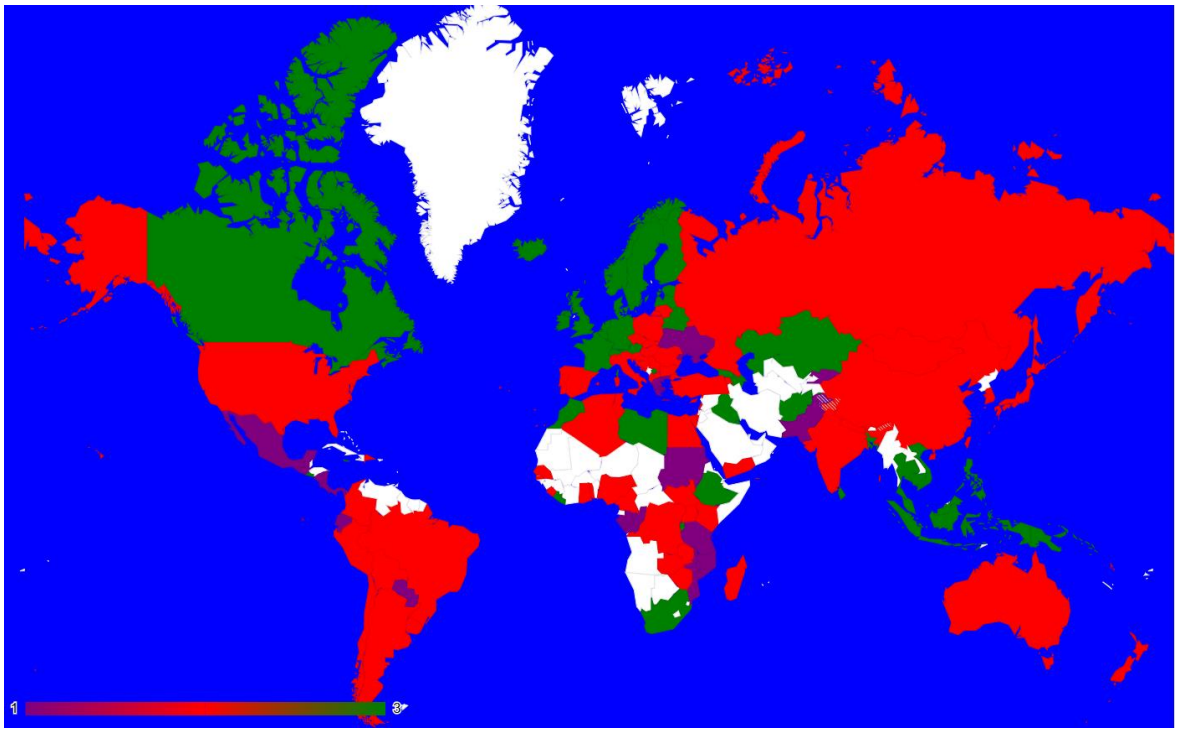

Figura 64; Clasificación año 2010

Año 2013

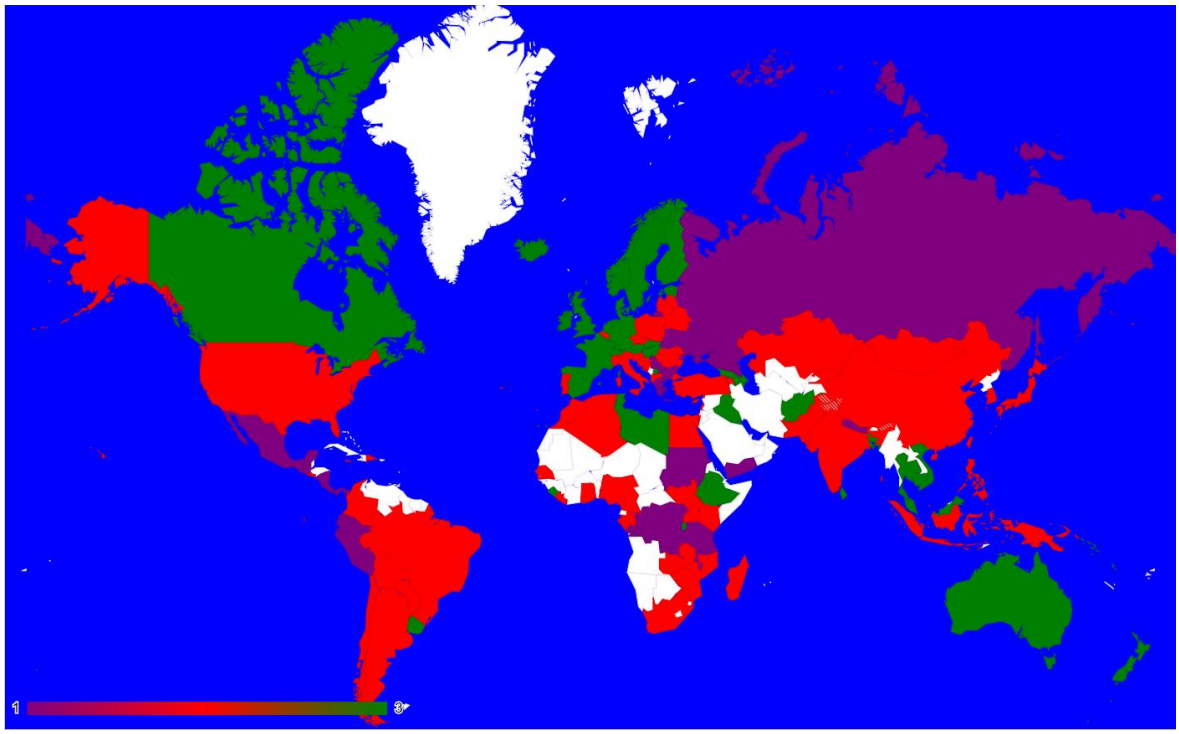

Figura 65: Clasificación año 2013

Año 2016

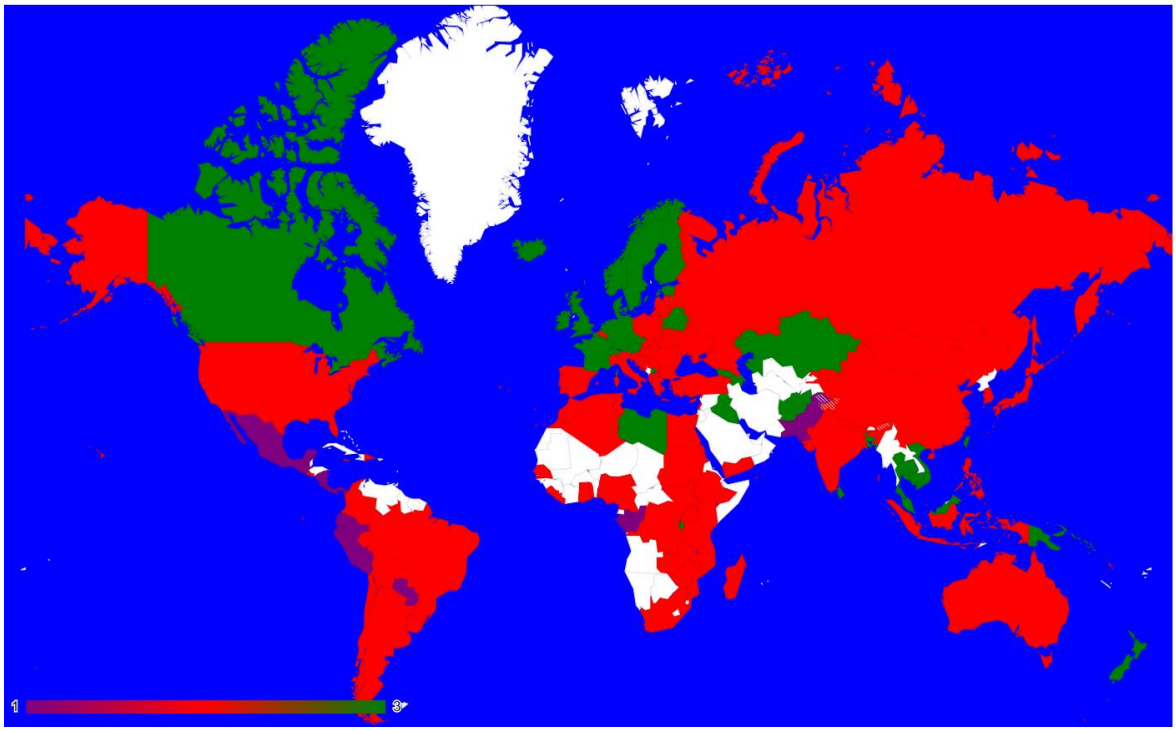

Figura 66: Clasificación año 2016

Año 2017

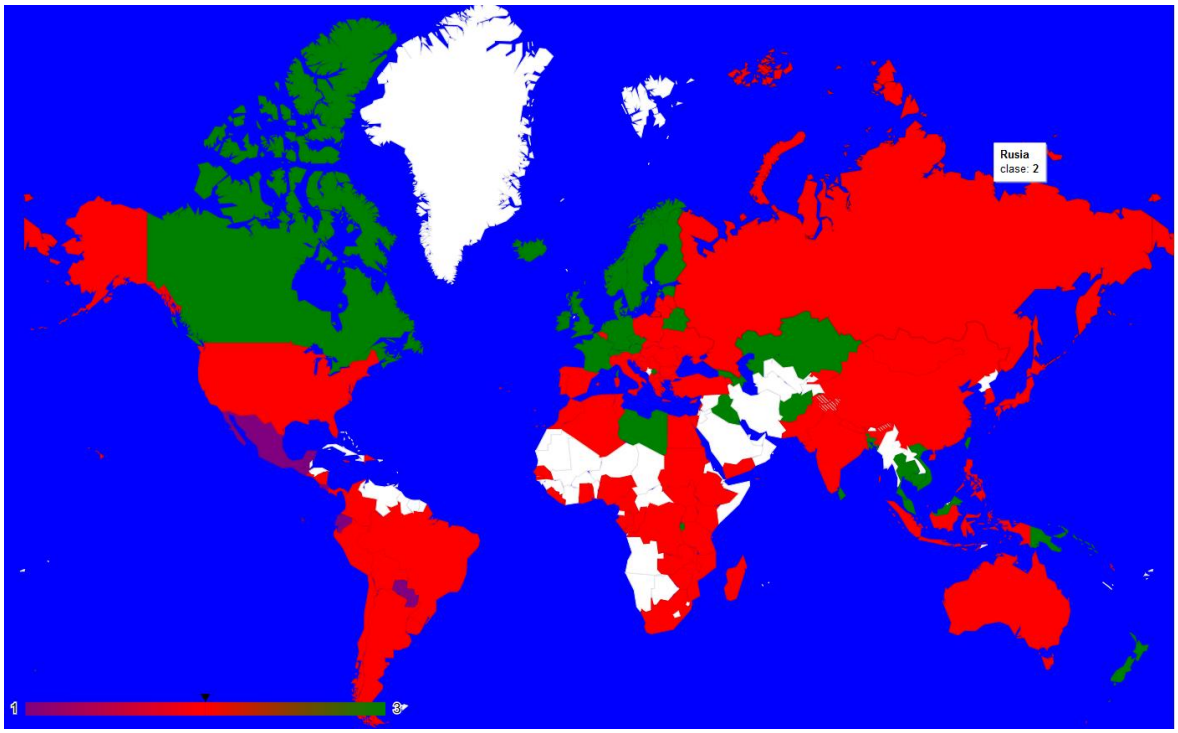

Figura 67: Clasificación año 2017

Año 2018

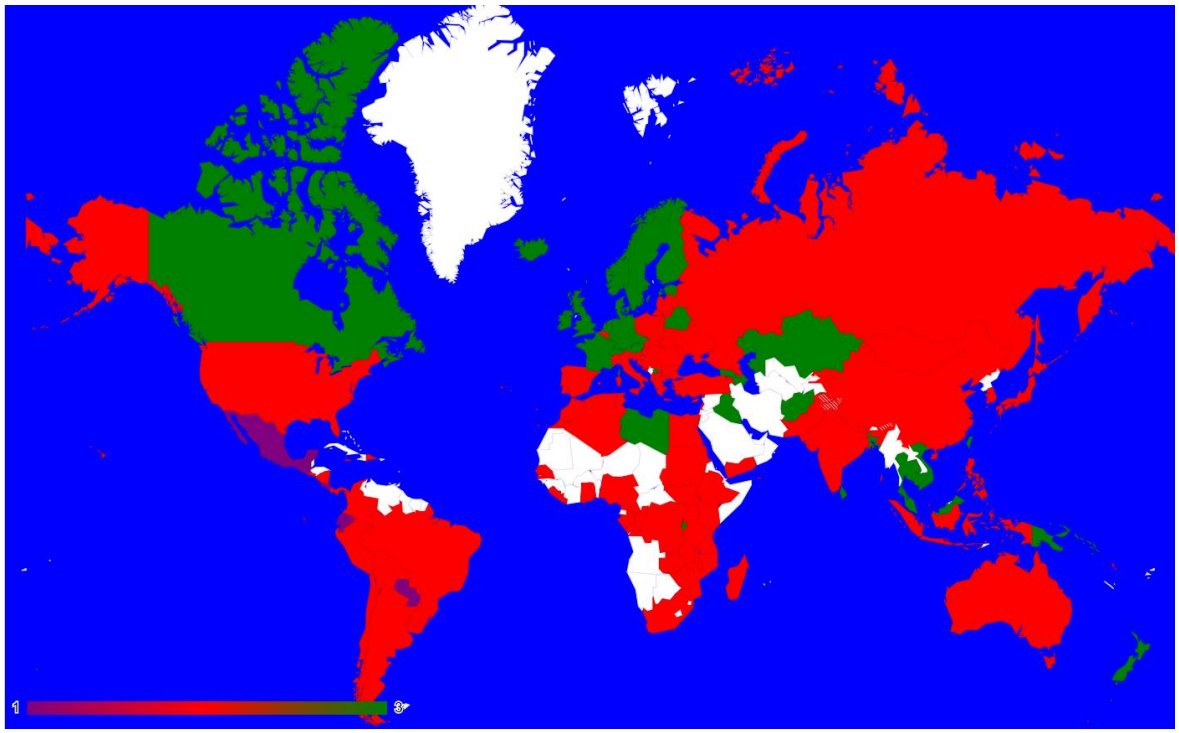

Figura 68: Clasificación año 2018

Año 2019

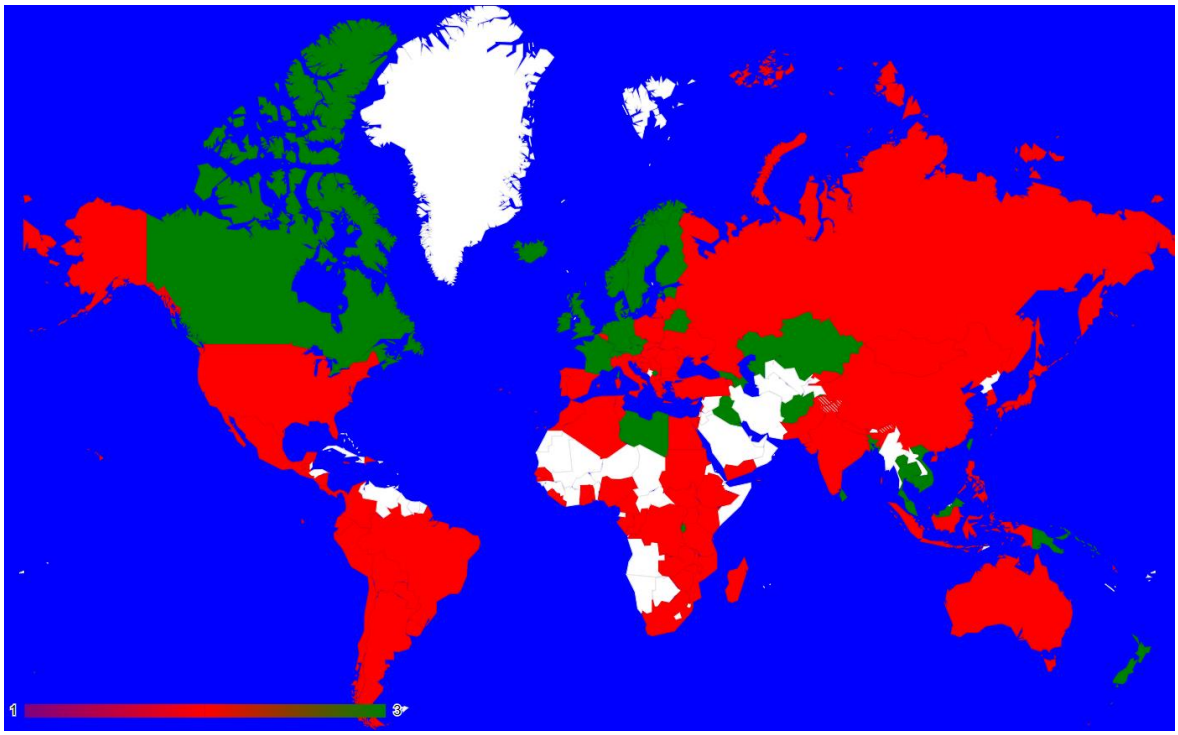

Figura 69: Clasificación año 2019

Año 2020

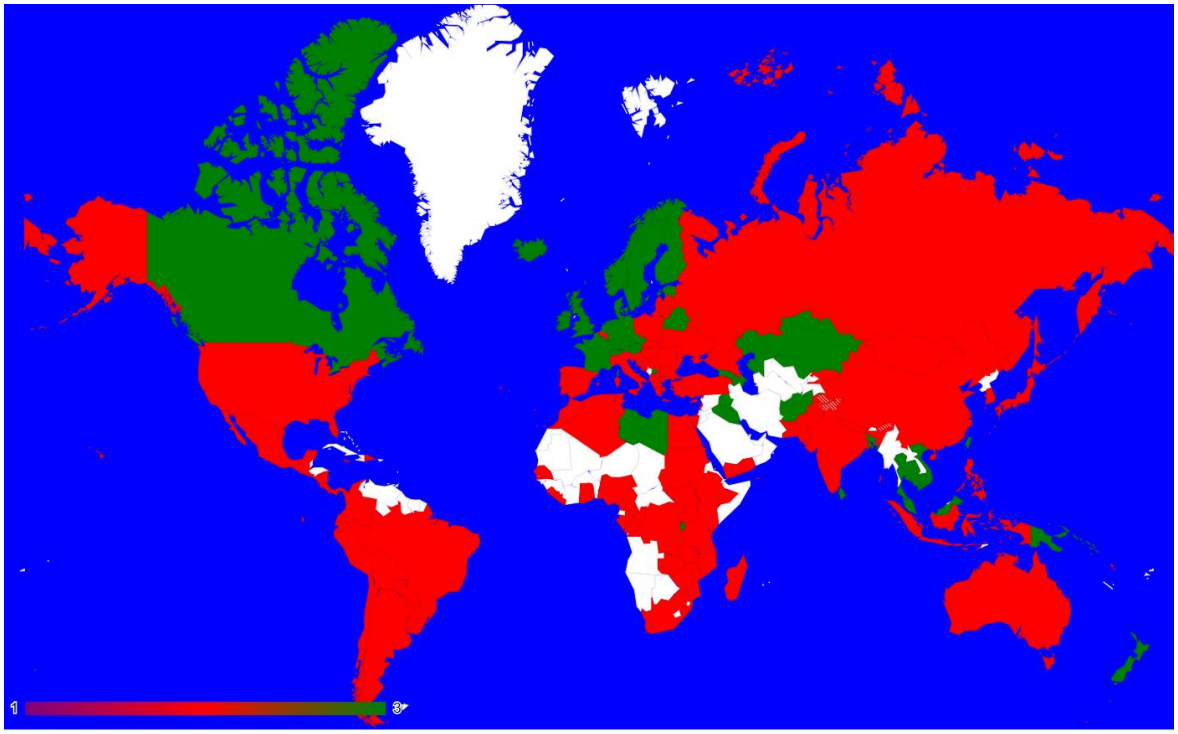

Figura 70: Clasificación año 2020

6 Bibliografía

[1] CENTRO DE INVESTIGACIONES SOCILÓGOCIAS (2015): Estudio nº 3121. Barómetro de diciembre 2015.Página 6

[2] Índice de Percepción de la Corrupción 2015 de Transparency International. Fecha de consulta 5 de Septiembre de 2016. Disponible en http://transparencia.org.es/wp-content/uploads/2016/01/tabla_sintetica_ipc-2015.pdf

[3] Entropía Diccionario de la Lengua Española RAE (http://dle.rae.es/?id=FpmDaOB) Fecha de consulta: 30 de Diciembre de 2016

[4] https://es.wikipedia.org/wiki/Teor%C3%ADa_de_la_informaci%C3%B3n . Fecha de consulta: 30 de Diciembre de 2016

[5] Information Theory for Intelligent People. Simon DeDeo Página 2. Disponible en http://santafe.edu/~simon/it.pdf.

[6] http://delta.cs.cinvestav.mx/%7Emcintosh/oldweb/s1998/abdiel/node3.html. Fecha de consulta 30 de Diciembre de 2016

[7] http://transparencia.org.es/ Visitado 30 de Diciembre de 2016

[8] https://en.wikipedia.org/wiki/Weka_(machine_learning)#External_links Fecha de consulta 2 de Enero de 2017

[9] https://en.wikipedia.org/wiki/Timeline_of_machine_learning Fecha de consulta 2 de Enero de 2017

[10] http://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#6de60314323f Fecha de consulta 2 de Enero de 2017

[11] Corrupción Diccionario de la Lengua Española RAE http://dle.rae.es/?id=V3cEQxK Fecha de consulta 10 de Enero de 2017